

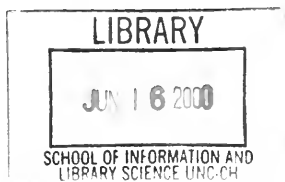
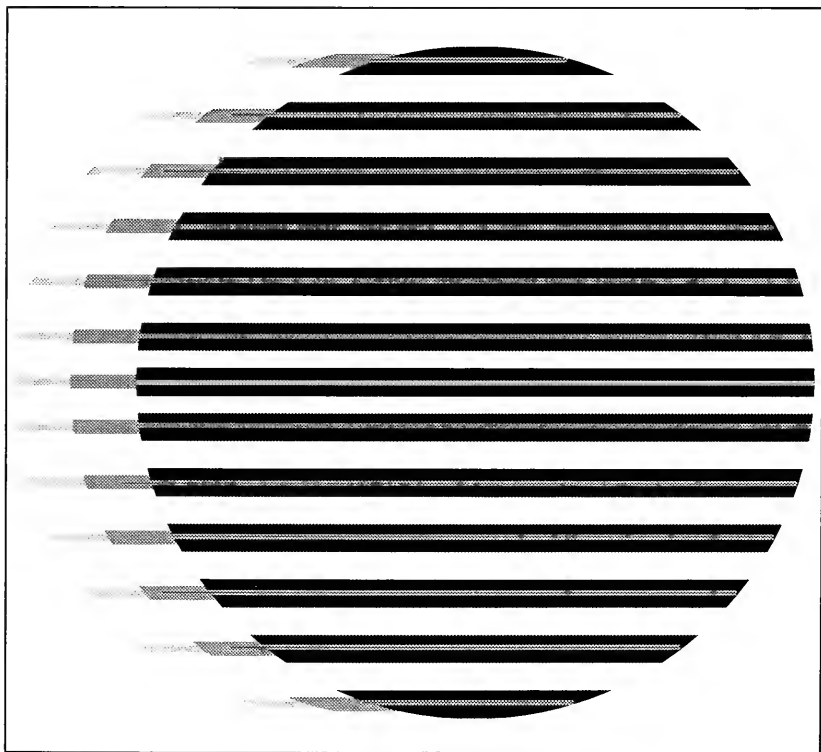
# IASSIST

Q U A R T E R L Y

VOLUME 21

Fall 1997

NUMBER 3



Digitized by the Internet Archive  
in 2010 with funding from  
University of North Carolina at Chapel Hill

Printed in the USA

<http://www.archive.org/details/iassistquarterly213inte>

# IASSIST QUARTERLY

The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

## Information for Authors:

The QUARTERLY is published four times per year. Authors are encouraged to submit papers as word processing files. Hard copy submissions may be required in some instances. Word processing files may be sent via email to [jtstratford@ucdavis.edu](mailto:jtstratford@ucdavis.edu). Manuscripts should be sent to Editor: Juri Stratford, Government Information and Maps Department, Shields Library, University of California, 100 North West Quad, Davis, California 95616-5292. Phone: (530) 752-1624.

The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science* 30(2):77-82, March 1979. Announcements of conferences, training sessions, or the like, are welcomed and should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event.

## Editors

**Karsten Boye Rasmussen,**  
Eckersbergsvej 56,  
5230 Odense M,  
Denmark.  
Phone: +45 6612 9811,  
Email: [boye@get2net.dk](mailto:boye@get2net.dk)

**Juri Stratford**  
Government Information and  
Maps Department,  
Shields Library,  
University of California,  
100 North West Quad,  
Davis, California 95616-5292  
Phone: (530) 752-1624.  
Email: [jtstratford@ucdavis.edu](mailto:jtstratford@ucdavis.edu)

## Production

**Laura Bartolo,**  
Libraries and Media  
Services,  
Kent State University,  
Ohio 44242.  
Phone: (330) 672-3024, x31.  
Email:  
[lbartolo@kentvm.kent.edu](mailto:lbartolo@kentvm.kent.edu)

**Walter Piovesan**  
Research Data Library  
Simon Fraser University  
Burnaby, B.C.  
Canada V5A 1S6.  
Phone: (604) 291-5937.  
Email: [walter@sfu.ca](mailto:walter@sfu.ca)

Title: Newsletter - International Association for Social  
Science Information Service and Technology

ISSN - United States: 0739-1137 © 1997 by IASSIST. All  
rights reserved.

# C O N T E N T S

Volume 21

Number 3

Fall 1997



## FEATURES

- 4 Changing the Way the United States  
Measures Income and Poverty: A Progress  
Report  
*Daniel H. Weinberg & Charles T. Nelson*
- 22 Theoretical and Technical Solutions for  
Preservation of Electronic Records in  
Finland  
*Matti Pulkkinen*
- 25 Tying Everything Together with a  
Relational Database  
*Pat Hildebrand*
- 28 Emerging Internet Image Archives  
Visualizing Biological Species and Medical  
Conditions.  
*Leslie A. Brownrigg*
- 32 Information that Come as Images:  
Overview of Issues  
*Repeke de Vries*
- 36 Categorizing Event Sequences Using Regular  
Expressions  
*Lisa Sanfilippo & John Van Voorhis*

# Changing the Way the United States Measures Income and Poverty: A Progress Report<sup>1</sup>

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect the views of the Census Bureau or the U.S. government. The authors would like to acknowledge and thank the following people for their comments and

suggestions—Nancy Gordon, Edward Welniak and the coauthors of the other papers cited in this report; they bear no responsibility for any errors that remain.

by *Daniel H. Weinberg &  
Charles T. Nelson*

questionnaire was expanded to identify over 50 sources of income and recording of up to 27 different income amounts, including receipt of numerous noncash benefits, such as food stamps (coupons used as cash for qualified food purchases), and housing assistance. Except for minor wording changes, those questions are still in use today. The survey was converted to a computer-assisted interviewing mode in 1994.

## I. BACKGROUND — THE OFFICIAL DEFINITION

The United States Census Bureau has been compiling income estimates annually since 1947. These estimates are from the Current Population Survey (CPS), a nationwide random sample of households, whose primary purpose is to collect labor force information monthly. In March of each year (April prior to 1956), data are collected on the household's income for the previous calendar year.

The official definition of income is not specified in law or regulation. In effect, what is included in income depends on the questions asked. As survey researchers know, the more questions one asks about income by source, the better able respondents are to identify all income. Initially, there were only two questions asked of each adult:<sup>2</sup> (1) "How much did ... earn in wages and salaries in 1947?" and (2) "How much income from all sources did ... receive in 1947?". In 1949, self-employment income was asked separately and in 1950 farm and nonfarm self-employment income was asked separately. In 1962, the Census Bureau began systematically assigning values to missing income items (based on reported characteristics using the "hot deck" method). In March 1967, the number of income questions was again expanded, from four to eight categories. These additional items dealt with Social Security, interest, dividends, and rent. In 1968, interest, dividends, rents, and royalties were combined into one question and separate questions were added on public assistance and on unemployment and workers' compensation. In 1975, the number of income questions increased from eight to eleven through addition of a question on the Supplemental Security Income program, a question on Aid to Families with Dependent Children and general assistance, and private and government pension income. A major change took place in 1980 — the

The data on income thus cover money income received (exclusive of certain money receipts such as capital gains) before payments for items such as personal income taxes, Social Security payroll taxes, and union dues. Money income does not reflect the fact that some families receive part of their income in the form of noncash benefits, such as food stamps, health benefits, rent-free or subsidized housing, and goods produced and consumed on the farm. In addition, money income does not reflect the fact that noncash benefits are also received by some as fringe benefits, e.g. the use of company cars, and full or partial payments by business for retirement programs, medical insurance, and educational expenses.

Moreover, for many different reasons, there is a tendency in household surveys for respondents to underreport their income. From an analysis of independently derived income estimates, it has been determined that income earned from wages or salaries is much better reported than other sources of income and is nearly equal to independent estimates of aggregate earnings (Coder and Scoon-Rogers, 1996). Among the least well-reported sources are interest and dividends. The detailed components of money income are presented in the Appendix.

## II. ALTERNATIVE MEASURES OF INCOME

Because money income is but one measure of economic well-being, the Census Bureau also reports on 14 other definitions of income (the series begins in 1979). While not exhaustive, they do illustrate different perspectives on what could be included.

*Definition 1.* Money income excluding capital gains before taxes. This is the official definition described above.

*Definition 2.* Definition 1 less government cash transfers. Government cash transfers include nonmeans-tested transfers such as Social Security payments, unemployment compensation, and government educational assistance (e.g., Pell Grants), as well as means-tested transfers such as Aid to Families with Dependent Children (AFDC), Temporary Assistance to Needy Families, and Supplemental Security Income (SSI).

*Definition 3.* Definition 2 plus capital gains. Realized capital gains and losses are simulated as part of the Census Bureau's Federal individual income tax estimation procedure. While the Census Bureau has access to some income information on individual tax returns that can be matched (with substantial time lag) to survey data, actual capital gains or losses or tax liability are not known.

*Definition 4.* Definition 3 plus imputed health insurance supplements to wage or salary income. Employer-paid health insurance coverage is treated as part of total worker compensation; no other benefits paid for or provided by employers are estimated.

*Definition 5.* Definition 4 less payroll taxes. Payroll taxes are payments for Social Security old age, survivors, and disability insurance, and for hospital insurance (Medicare).

*Definition 6.* Definition 5 less Federal income taxes. The effect of the Earned Income Tax Credit, targeted to low-income workers, is shown separately in Definition 7.

*Definition 7.* Definition 6 plus the Earned Income Tax Credit.

*Definition 8.* Definition 7 less state income taxes.

*Definition 9.* Definition 8 plus nonmeans-tested government cash transfers. Nonmeans-tested government cash transfers include Social Security payments, unemployment compensation, workers' compensation, nonmeans-tested veterans' payments, U.S. railroad retirement, Black lung payments, and Pell Grants and other government educational assistance. (Pell Grants are income-tested but are included here because they are very different from the assistance programs included in the means-tested category.)

*Definition 10.* Definition 9 plus the value of Medicare. Medicare is counted at its fungible value.<sup>3</sup>

*Definition 11.* Definition 10 plus the value of regular-price school lunches.

*Definition 12.* Definition 11 plus means-tested government cash transfers. Means-tested government cash transfers include AFDC, SSI, other public assistance programs, and means-tested veterans' payments.

*Definition 13.* Definition 12 plus the value of Medicaid. Medicaid is counted at its fungible value.

*Definition 14.* Definition 13 plus the value of other means-tested government noncash transfers. Including food stamps, rent subsidies, and free and reduced-price school lunches.

*Definition 15.* Definition 14 plus net imputed return on equity in one's own home. This definition includes the estimated annual benefit of converting one's home equity into an annuity, net of property taxes.

Table 12 is a reproduction of a table from U.S. Bureau of the Census (1996a) illustrating the different distributions of income that these definitions imply.<sup>4</sup> Table 5 (U.S. Bureau of the Census, 1996b) illustrates this effect on poverty estimates.

These alternative definitions illustrate the dilemma faced by official statisticians when presenting income statistics. Different definitions serve different purposes. Money income has its uses — it represents command over the resources available to purchase the necessities of life in the open market, including meeting the obligations of citizenship (taxes). Definition 4 probably comes closest to measuring what resources would be available in the absence of government, except that some benefits paid for or provided by employers are not included and others are mandated by the government, some benefits are not provided by employers because they are provided by the government, and work effort is presumably reduced by the existence of a tax on earnings. Definition 8 is closest to after-tax income. Disposable income tries to take account of the effect of taxes and transfers on the household's command of resources — definition 14 probably comes closest to that approach. Finally, in definition 15 there is an attempt to include the income equivalent value of owning one's own home in that such an asset reduces the need for additional expenditures on shelter.

### III. CONSIDERATIONS IN MEASURING POVERTY

Formal measurement of poverty in the United States is less than three decades old. Not since the adoption of official poverty thresholds by the Federal government in the late 1960's has there been such a great interest as now in examining and possibly respecifying the thresholds and the income compared with them. The official poverty thresholds in use today by the U.S. Bureau of the Census to measure poverty have their basis in work by Orshansky

(1963, 1965). Orshansky started with a set of minimally adequate food budgets calculated for families of various sizes and composition by the U.S. Department of Agriculture for 1961. Based on evidence from the 1955 Household Food Consumption Survey, she determined that expenditures on food represented about one-third of after-tax income for the typical family. This relationship yielded a “multiplier” of three, that is, the minimally adequate food budgets were multiplied by a factor of three to obtain 124 poverty thresholds that differed by family size, number of children, age and sex of head, and farm or nonfarm residence (ad hoc adjustments were made for families of size one and two).

In 1969, the U.S. Bureau of the Budget (now the U.S. Office of Management and Budget — OMB) adopted the Orshansky measure using pre-tax income as the standard government poverty measure, mandating that thresholds be adjusted for inflation using the Consumer Price Index (CPI) published by the U.S. Bureau of Labor Statistics. With only minor modifications since then (mostly reducing the number of categories, now 48), the Orshansky thresholds still form the basis for the official poverty statistics.<sup>5</sup>

When considering the adequacy of the official poverty thresholds, it is critical to realize that one cannot separate the issue of income measurement from poverty definition. When one defines the level of resources needed to be non-poor, one must also determine which resources are to be counted. Therefore, the discussion below covers both income measurement and poverty definition issues; income measurement is discussed first.<sup>6</sup>

Whatever poverty thresholds are chosen should be the result of a carefully specified process that cannot be changed arbitrarily from year-to-year, and should be capable of being updated at reasonable intervals as the economic circumstances of the society and the behavior of its demographic and economic components change.

#### A. DEFINING INCOME FOR MEASURING POVERTY

The key measurement issues are three — valuing and counting noncash income, subtracting taxes, and reducing survey underreporting and nonsampling errors. Also of interest is whether to continue to publish official estimates based on the CPS or switch to a newer survey designed to collect better income information, the Survey of Income and Program Participation (SIPP).

##### A.1. Noncash income

The issue of valuing noncash income spans the income distribution. A more comprehensive income measure, such as definition 14 above, would place a value not only on noncash government transfers, such as food stamps, which typically go to low-income families, but also on elements of nonwage compensation (from employer-paid health

insurance to company cars) that typically go to earners at all income levels or only at high levels. The noncash income of U.S. families has grown substantially in the past 25 years. In the 1990’s, over half of government transfer spending for the poor is in the form of noncash benefits (U.S. Bureau of the Census, 1996a), whereas the only noncash benefit program that predated the 1960’s “War on Poverty” was subsidized (public) housing. This growth of benefits to the poor has been paralleled by a growth of nonwage compensation to wage earners, induced in part by tax laws exempting such compensation from income and payroll taxes, and by growth in health benefits for the elderly. By 1996, employer costs for nonwage compensation had grown to over one-quarter (28.4 percent) of total compensation costs, up from 19.4 percent in 1966.<sup>7</sup> Further, nearly two-thirds of households own homes, which provide them with additional noncash income in the form of housing services.

Of key concern to understanding the well-being of U.S. households is the valuation of medical benefits, both the government health programs—Medicare (medical aid to the elderly and severely disabled) and Medicaid (medical aid to a portion of the poor)—and employer-paid health insurance. The valuation of medical benefits is particularly difficult since coverage of high medical expenses for people who are sick does nothing to improve their poverty status (although the benefits clearly make them better off). Even if one imputes the value of an equivalent insurance policy to program participants, these benefits (high in market value due to large medical costs for the fraction who do get sick), and cannot be used by the recipients to meet other needs of daily living. Accordingly, the Census Bureau developed a not-altogether-satisfactory method, termed fungible value (described in footnote 2), to avoid giving too high a value of these benefits to those toward the low end of the income scale. Note that this is not a problem for countries with universal health care systems.

##### A.2. Disposable income

Even though Orshansky’s original calculations were based on post-tax income, poverty has always been calculated for the official statistics using pre-tax income because of the limited information collected on the CPS. After-tax income is a better measure of the ability to meet the daily necessities of life than is money income. Also important, in calculating disposable income though, is to address the advisability of deducting work expenses for wage earners such as child care, uniforms, and transportation costs.

##### A.3. Other issues

As noted earlier, research matching household survey responses to Federal income tax returns and comparing them with national income accounts has revealed substantial areas where the level and receipt of certain

income sources is underreported. Attempts to reduce underreporting were made by revising the language used in the CPS questionnaire (and using a shorter reference period) when the SIPP was launched. This was only partially successful, and response errors remain.

While current procedures of the Census Bureau reweight the data for full interview nonresponse and impute appropriate income responses for individual unanswered questions (item nonresponse), these corrections are insufficient to fully resolve the problem. Procedures to enhance the data through microsimulation or other means are being investigated, along with continued improvement in imputation for nonresponse.

In most societies, "underground," "nonmarket," or "black market" income from legal or illegal activities is typically poorly reported by household respondents to government surveys (or not even collected) and consequently is substantially omitted from official income statistics. This income ranges from barter transactions to home production (e.g., home gardens) to illegal income. Researchers are a long way from measuring this activity accurately, however, so including this income in official statistics would be quite difficult.

It has been suggested that consumption is a better measure of well-being than income (see Cutler and Katz, 1991, and Slesnick, 1993). If a family can maintain its consumption through judicious use of assets when income falls, is it truly poor? Unfortunately, it is difficult to collect accurate annual data on consumption or even expenditures. Further, consumption reflects choices on how to allocate resources, rather than need. Nevertheless, fuller investigation of a consumption-based measure would be useful.

A final issue of income measurement is the choice of surveys to use. As mentioned briefly above, the SIPP questionnaire design, as crafted to reduce income underreporting, does succeed for almost all income sources.<sup>8</sup> Yet, when compared with the CPS, it has historically had several drawbacks—a smaller sample size (one-third as large) and necessarily slower data release because of its much greater complexity. These defects are compensated for by the SIPP having greater income detail, both in number of sources and in time segments (by having monthly as opposed to the CPS's annual statistics,) and lower underreporting. The new version of the SIPP, as implemented in 1996, increased the sample size substantially (to 36,700 households) and oversampled low-income households. National estimates from the SIPP will then be comparable to or better than (in terms of sampling error) those from the CPS (reduced to 48,000 households but inefficient for national estimates because it uses a state-based design). One drawback for obtaining a consistent time series of annual national income or poverty estimates from the SIPP, though, will be sample attrition and time-in-

sample bias as current plans call for only one SIPP panel to be in the field during any one four-year period. The CPS sample is constantly refreshed by new households.

While the timeliness issue may never be resolved fully in SIPP's favor, the SIPP can provide a preliminary estimate on much the same schedule as the CPS. Still, it is desirable to view the surveys complementarily. If modeling using administrative records can correct underreporting errors in both surveys, they would then give the same aggregate statistics. The CPS could be used for a quick snapshot, consistent with data collected since 1947 (the SIPP began in 1983), while the SIPP would be used for more detailed estimates, for subannual and multiyear estimates, and for understanding other dimensions of poverty (assets, disability, gross flows, and other dynamic aspects).<sup>9</sup>

## ***B. SETTING THRESHOLDS TO DEFINE POVERTY***

With an absolute measure of poverty, there are key decisions to be made about determining the appropriate level for poverty thresholds. The key research issues addressed here are minimal consumption levels for specific commodities, ways of correcting for differences in family size and composition, and ways of correcting for cost-of-living differences across time and among areas.

### ***B.1. Minimal consumption standards***

Minimal consumption standards for all necessary commodities could in theory be established, perhaps by an expert panel, but doing so would raise difficult ethical issues about which commodities to include (e.g., is a telephone a necessity?). One alternative is to define minimal consumption standards for a limited number of necessities (e.g. food, clothing, shelter) and obtain a poverty threshold by using a multiplier to account for necessities not measured.<sup>10</sup>

### ***B.2. Equivalence scales***

The relationship embodied in the current U.S. poverty thresholds among families of different sizes (termed the equivalence scale) is supposed to represent the different relative costs of supporting those families at a minimally adequate levels. In fact, the relationship is based solely on the relative food costs as they existed in 1961 and include some unfortunate anomalies (see Ruggles, 1990, pp. 64–68). While it is possible to develop minimal budgets for every type and size of family separately and thus eliminate the need for equivalence scales entirely, in practice it is difficult to do so. No one scale now exists that is generally accepted. Issues in developing equivalence scales include which distinctions in family circumstances (e.g. owner/renter) should lead to different thresholds, how resources are shared within the family or household, and whether a more useful basis for determining poverty is the household (those living in one housing unit) rather than the family

(those in one household related by blood or marriage). See Betson (1996) for a further discussion of these issues.

### B.3. Cost-of-living differences

In as large and diverse a country as the U.S., there are significant differences in the cost of living among localities. Unfortunately, there are no currently available data upon which to estimate interarea price differences reliably. (See Kokoski et al., 1992, and Moulton, 1992, for some work in this area.)

A related price issue is how to adjust for inflation. The U.S. poverty thresholds now use the CPI to adjust thresholds over time. If the measurement of minimal consumption is used as the basis for new thresholds, presumably this should be the basis every year, with components, prices, and multipliers reestimated as often. Clearly this is not practical. A reasonable compromise might be to respectify and reestimate the minimal consumption bundle at prespecified intervals as market baskets become outdated, say every ten years, and use the CPI for interim adjustments. The market basket used for the CPI itself is typically reviewed and respecified once every ten years or so.<sup>11</sup>

### C. *THE COMMITTEE ON NATIONAL STATISTICS REPORT*

The National Academy of Sciences' Committee on National Statistics (CNStat) released a report in May 1995 entitled Measuring Poverty: A New Approach (Citro and Michael, 1995). In that report, the committee recommended that the Federal government redefine the way it measures poverty. OMB has requested that experts from the Census Bureau and other agencies examine technical methods for doing so.

The key changes they recommend are threefold: change the income measure, change the poverty thresholds, and change the survey used. To change the income measure from the current money income definition, they propose to add noncash benefits, subtract taxes, subtract work expenses, subtract child care expenses, subtract child support paid, and subtract medical out-of-pocket expenses (MOOP). The poverty thresholds are to be based on food, clothing, shelter, and "a little bit more" (75-83% of median expenditures on these items multiplied by 1.15-1.25), a new equivalence scale, an allowance for geographic variation, and are to be updated annually based on growth in median expenditures. Finally, the panel recommended that the government use the SIPP instead of the March CPS to collect the basic income and poverty-related data.

Among the technical issues to be resolved before implementing such a new measure are the following:

1. Reestimating the valuation methodologies for

government noncash transfer programs including school lunches, food stamps, and housing benefits; developing new estimation methodologies for additional programs and possibly developing a new methodology for valuing Medicare and Medicaid (depending on whether the subtraction of MOOP is adopted or not);

2. Completing development of a tax simulation model for SIPP;
3. Developing a methodology for estimating MOOP (e.g. a statistical match of the National Medical Expenditures Survey to SIPP) or reestimation of employer contributions to health insurance using more recent data;
4. Estimating and imputing work and child care expenses;
5. Redesigning the SIPP sampling scheme to maximize reliability of a time series of cross-section estimates while maintaining some longitudinal estimation capabilities, taking account of the need for state-level estimates, and minimizing the attrition bias;
6. Reviewing the Consumer Expenditure Survey to improve its effectiveness for its new dual role (defining the market basket for the Consumer Price Index and the poverty thresholds) and possibly preparing for consumption-based rather than income-based poverty estimates in the future;
7. Creating a time series of poverty estimates from the SIPP and developing methods to impute additional variables to the CPS to develop comparable time-series data for that survey;
8. Doing substantial further work on income underreporting and imputation models;
9. Adding child support and alimony paid questions to CPS; and
10. Developing and adding "medical care risk" and possibly medical expenditures questions to SIPP to supplement the poverty measure if medical care costs and benefits are excluded from the measure.

Even if these technical issues can be resolved expeditiously, there are still policy issues that must be debated and resolved before a new measure is adopted. These include:

1. *Including or excluding medical costs and benefits.*  
On the one hand, the CNStat recommended excluding MOOP, employer contributions to health insurance, and benefits from medical transfer programs from income. On the other hand, adopting as official the current (experimental) practice of including them would require

improving the current method for valuing medical transfer program benefits, measuring medical needs more accurately, and updating the methodology for imputing employer contributions to health insurance.

*2. Basing thresholds on a pre-specified fraction of median expenditures.* How might the public and Congress react to a new poverty threshold that showed millions more poor persons than the current measure? Are we confident enough about the quality of (i.e. lack of biases in) the Consumer Expenditure Survey data to use it as the arbiter of the poverty level? It may be that the likely acceptance of any new definition would be enhanced if the new index were “chained” to the old by matching the overall poverty rate obtained (but allowing the distribution to vary).

*3. Developing geographical cost-of-living variations.* It is clear that the cost of living differs substantially from place to place, and different choices of methodology to reflect this fact would have different implications. If geographic variation is to be incorporated, some method for periodically updating the thresholds for relative price changes among areas would also need to be established.

*4. Annual inflation updating.* The panel proposed using the rate of growth in expenditures to index the thresholds. This is an attempt to introduce some deliberate “relativity” into the measure and would have quite different ramifications from using the Consumer Price Index.

*5. Choosing the equivalence scale.* Choice of the scale would inevitably alter the distribution of the poor.

*6. Underreporting.* If the technical issues about how to do so are resolved, should the income statistics from the survey be adjusted for underreporting based on administrative data and modeling?

*7. Review and Revision.* Should any new definition include a regular cycle of review and revision based on pre-specified criteria (CNStat recommended once a decade)?

Open debate of these issues seems the most likely way to resolve them, potentially leading to a new way of measuring poverty that OMB would approve and that other policy makers would accept as an improved methodology for measuring poverty in the United States.

#### ***D. CENSUS BUREAU POVERTY REDEFINITION RESEARCH***

In order to provide a basis on which some of these issues can be resolved, the Census Bureau and other U.S. government agencies have begun research studies.

#### ***D.1. Census Bureau-Bureau of Labor Statistics Study***

The CNStat report on redefining poverty contained sweeping recommendations for changing the way poverty is defined in the U.S. Recent joint research by the Bureau of the Census and the Bureau of Labor Statistics (BLS) (Garner et al., 1997) examined two of these issues — changing the income definition and modifying the poverty thresholds.

In formulating poverty thresholds, BLS researchers started by implementing the basic recommendations from the CNStat report. Some of the CNStat panel recommendations regarding thresholds were given as ranges. Thus, some simplifying assumptions were made. First, the panel recommended a range of thresholds, with a lower bound based on 78 percent of median expenditures for food, clothing, and shelter and a multiplier of 1.15 to account for other needs. The upper bound was based on 83 percent of the median and a multiplier of 1.25. In the Garner et al. paper the midpoint of this range was used. The other simplifying assumption was for the equivalence scale (the relationship between thresholds for different family sizes). The panel recommended a range of economy scale factors of 0.65 to 0.75 and again they choose the midpoint — 0.70. Thresholds were computed for the years 1990 through 1995.

On the resource side, the panel’s recommendations were followed to the extent possible. The only recommendation not followed (because of a lack of data) was their recommendation to subtract child support paid from income when computing a poverty resource measure. Though the panel recommended changing the official source of poverty statistics in the U.S. from the CPS to the SIPP, the initial work was based on the CPS. At this time, the CPS is the only survey with a working tax simulation model and in-kind benefit valuation procedures, both necessary ingredients for producing a resource measure based on the panel’s recommendations.

The report found that the threshold computation methods as recommended by the panel result in relatively stable thresholds over time (at least over the 1990-1995 period measured in this study), and the resulting poverty rates based on applying the panel’s basic resource definition to these thresholds also showed relatively stable results. In fact, though the panel’s recommendations result in significantly higher poverty rates than the U.S. official estimates, the trends based on the official estimates and the panel’s recommended method show very similar trends over the 1990-1995 period (see Figure 1). Differences across subgroups were also found to be stable over time. However, the key change under the proposed definition of poverty is in the composition of the poverty population. Consistent with the panel’s findings, poverty rates under the recommended poverty measure are significantly higher among groups with relatively low official poverty rates (for

example, Whites or those living in married-couple families). Groups with relatively high poverty rates, on the other hand, did not tend to have very different poverty rates under the revised measure. Thus, an effect of moving to the recommended poverty measure would be to narrow the gaps that now exist in the U. S. between high- and low-poverty groups (married-couple and single-parent families, Whites and Blacks, etc.). Put another way, under the revised measure, the poverty population looks more like the total population in terms of demographic and socioeconomic characteristics. (See Table 1 and Figures 2-4.)

Other, slightly different poverty thresholds were also examined in the Census-BLS study. One modification, which was suggested by the panel, was to define shelter costs by their rental equivalent value. This technique resulted in higher poverty thresholds (and higher poverty rates), and appeared to have some effect on the composition of the poverty population (further narrowing the gaps, for example, between high- and low-poverty groups). Another set of thresholds was based on alternative multipliers that were computed more precisely than those used in the Panel's report. This modification resulted in little change in the composition of the poverty population.

#### D.2. Other Census Bureau Poverty Research

The panel recommended changing the source of official U.S. poverty estimates from the March CPS to the SIPP. As noted earlier, the SIPP is a longitudinal survey with: 1) a more detailed set of questions than the CPS, 2) a shorter reference period (4 months versus 12 months for the CPS), and 3) increased flexibility sufficient to add the questions required to measure poverty based on the broadened resource definition recommended by the panel. Questions have already been added to SIPP to collect some of this additional information, and a sample design change, in order to make SIPP a better cross-sectional survey (a requirement for measuring annual poverty changes) has been proposed, though not yet adopted.

The Census Bureau has also examined the panel's recommendations on work-related and child-care expenses (the panel recommended subtracting these costs from income when computing the poverty resource measure and has suggested alternative methods for imputing such costs). This research showed that using a definition of resources that excludes child care and other work-related expenses has a significant effect on poverty rates. In both CPS and SIPP-based analyses, the effect of using a resource definition that excluded these expenses was to raise children's poverty rates by about 3 percentage points. (See Short et al., 1996.)

Another area of research at the Census Bureau is on the housing subsidy valuation method. The value of public or subsidized housing is included in the recommended poverty

measure, and the current Census Bureau method for imputing such subsidies (on the CPS) is badly outdated. Current methods are being reviewed, and ways to implement this imputation on SIPP are being explored. A paper is planned for presentation in August (Eller and Naifeh, forthcoming).

The one major element of the panel's recommended resource measure not included in the Census Bureau-BLS study was the subtraction of child support paid, since this information was not available in the CPS. Data from SIPP indicate that the inclusion of such payments would increase the poverty rate by 0.3 to 0.5. Questions were added to the April 1996 CPS Supplement on child support to examine the feasibility of capturing this information on a regular basis on the March CPS. Data on child support paid are regularly collected on SIPP.

As already noted, the treatment of medical benefits and expenditures in defining poverty is a difficult one. Staff are currently examining the treatment of medical out-of-pocket expenditures in the definition of poverty (see Doyle, forthcoming(a)). To come up with a definition of income that excludes these expenditures, our current thinking is that statistically matching SIPP to another Federal government survey that includes detailed information about these expenditures (the Medical Expenditure Panel Survey) holds the most promise. In addition, staff are working on a proposed medical care risk index to complement the new poverty measure (to address another recommendation of the panel). (See Doyle, forthcoming(b).)

Since the panel recommended an after-tax income definition for its poverty measure, one problem with transferring the official poverty measure from the CPS to SIPP is the lack of a working tax simulation model based on the SIPP (since the early 1980's, the CPS has employed a model to estimate taxes). The Census Bureau, along with several other Federal agencies, supported the development of a SIPP-based tax model, and we are now in the process of exploring how to best incorporate this model into the Census Bureau's processing system.

Equivalence scales are an important issue in the formulation of poverty thresholds. Betson (1996) provides compelling evidence that the choice of equivalence scales has a significant effect on the composition of the poverty population. He also pointed to the need for continued research in this area.

In another paper, Betson (1995) examined the issue of home ownership and whether the flow of housing services from owner-occupied homes should be taken into account when defining poverty status. He found that counting the value of housing services would change the distribution of the poor, primarily by counting fewer of the elderly as poor.

Table 5. Poverty Rates: Official and Experimental by Race, Hispanic Origin, Family Type and Age: 1992

	Official	Experimental	Percent Difference
All Persons	14.8	19.9	34.5
White	11.9	17.1	43.7
Black	33.4	37.1	11.1
Hispanic Origin (of any race)	29.6	41.5	40.2
Married Couple	7.7	13.7	77.9
Female Household	39.0	42.8	9.7
Under 18 Years Old	22.4	27.1	21.0
18 - 64 Years Old	11.9	16.3	37.0
65 Years Old and Over	12.9	22.5	74.4

### E. CONCLUDING REMARKS

We believe that prospects for developing a consensus around a new measure of poverty in the United States are the highest since the current measure was adopted in the 1960's. Converting the measure to the SIPP is not costless, though, and budgetary pressures may cause a delay even if a broad methodological consensus is reached. Furthermore, delicate negotiations over broad policy issues must ensue before any change is made.

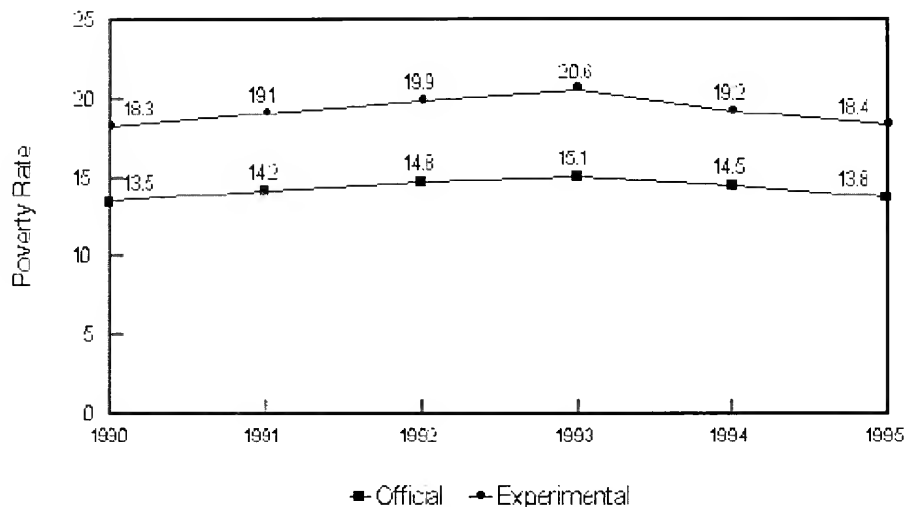
Readers are welcome to follow further developments as they happen. Visit the special poverty measurement web site at <http://www.census.gov/hhes/www/povmeas.html>.

### APPENDIX: DEFINITION OF MONEY INCOME

The current official U.S. definition of income is based on questions which are asked of each person in the CPS sample household 15 years old and over.<sup>12</sup> These questions cover the amount of money income received in the preceding calendar year from each of the following sources.

*Earnings from longest job (or self-employment) and other employment earnings* can be classified into three types: (1) Money wage or salary income is the total received for work performed as an employee during the income year. This category includes wages, salary, Armed Forces pay, commissions, tips, piece-rate payments, and cash bonuses earned, before deductions were made for items such as taxes, bonds, pensions, and union dues; (2) Net income from nonfarm self-employment is the net money income (gross receipts minus expenses) from one's own business, professional enterprise, or partnership. Gross receipts include the value of all goods sold and services rendered. Expenses include items such as costs of goods purchased, rent, heat, light, power, depreciation charges, wages and salaries paid, business taxes (not personal income taxes);<sup>13</sup> and (3) Net income from farm self-employment is the net money income (gross receipts minus operating expenses) from the operation of a farm by a person on their own account, as an owner, renter, or sharecropper. Gross receipts include the value of all products sold, payments from government farm programs, money received from the rental of farm equipment to others, rent received from farm property if payment is made based on a percent of crops produced and incidental receipts from the sale of items

# Figure 1. Poverty Rates: Official and Experimental: 1990-1995



such as wood, sand, and gravel. Operating expenses include items such as the cost of feed, fertilizer, seed, and other farming supplies; cash wages paid to farmhands; depreciation charges; cash rent; interest on farm mortgages; farm building repairs; and farm taxes (not state and Federal personal income taxes). The value of fuel, food, or other farm products used for family living is not included as part of net income.<sup>14</sup>

*Unemployment compensation* includes payments received from government unemployment agencies or private companies during periods of unemployment and any strike benefits received from union funds.

*Workers' compensation* includes payments received periodically from public or private insurance companies for injuries received at work.

*Social Security* includes Social Security (old age) pensions and survivors' benefits and permanent disability insurance payments

made by the Social Security Administration prior to deductions for medical insurance. Medicare reimbursements for health services are not included.

*Supplemental Security Income* includes payments made by

## Figure 2. Composition of the Poverty Population, Official and Experimental, by Race: 1992

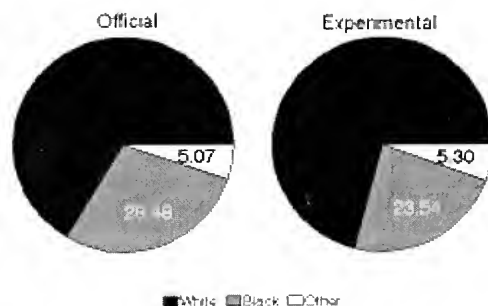
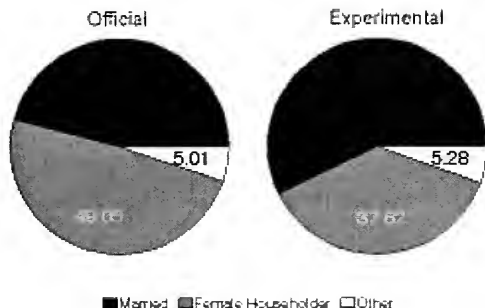


Figure 3. Composition of the Poverty Population, Official and Experimental, by Family Type. 1992



Federal, state, and local welfare agencies to low income persons who are 65 years old or over, blind, or disabled.

*Public assistance or welfare payments* include public assistance payments made to low-income persons, such as Aid to Families With Dependent Children, Temporary Assistance for Needy Families, and general assistance.

*Veterans' payments* include payments made periodically by the Department of Veterans Affairs to disabled members of the Armed Forces or to survivors of deceased veterans for education and on-the-job training, and means-tested assistance to veterans.

*Survivor benefits* include payments from survivors' or widows' pensions, estates, trusts, annuities, or any other types of survivor benefits. Payments can be reported from ten different sources: private companies or unions; Federal government (Civil Service); military; state or local governments; railroad retirement; workers' compensation; "Black lung" (miners') payments; estates and trusts; annuities or paid-up insurance policies; and other survivor payments.

*Disability benefits* include payments received as a result of a health problem or disability other than those from Social Security. Payments can be reported from ten sources: workers' compensation; companies or unions; Federal government (Civil Service); military; state or local governments; railroad retirement; accident or disability insurance; Black lung payments; state temporary sickness; or other disability payments.

*Pension or retirement income* includes payments reported from eight sources:

companies or unions; Federal government (Civil Service); military; state or local governments; railroad retirement; annuities or paid-up insurance policies; withdrawals from special (tax-favored) retirement accounts such as Individual Retirement Account (IRA's); or other retirement income.

*Interest income* includes payments received (or credited to bank accounts), from bonds, treasury notes, IRA's, certificates of deposit, interest-bearing savings and checking accounts, and all other investments that pay interest.

*Dividends* include income received from stock holdings and mutual fund shares. Capital gains from the sale of stock holdings are not included as income.

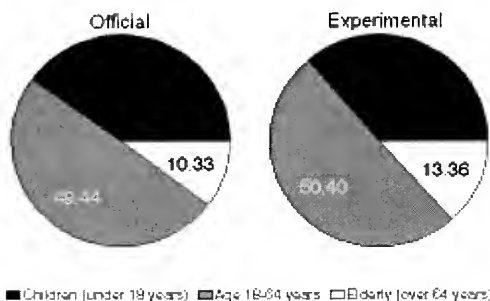
*Rents, royalties, and estates and trusts* include the net income from the rental of a house, store, or other property, receipts from boarders or lodgers, net royalty income, and periodic payments from estate or trust funds.

*Educational assistance* includes Pell Grants; other government educational assistance; any scholarships or grants; or financial assistance from employers, friends, or relatives not residing in the student's household.

*Child support* includes all periodic payments made by parents for the support of children, even if these payments are made through a state or local government office.<sup>15</sup>

*Alimony* includes all periodic payments to ex-spouses. One-time property settlements are not included.

Figure 4. Composition of the Poverty Population, Official and Experimental, by Age: 1992



*Financial assistance from outside of the household* includes periodic payments from nonhousehold members. Gifts or sporadic assistance is not included.

*Other income* includes all other regularly received payments that are not included elsewhere on the questionnaire. Some examples are state programs such as foster child payments, military family allotments, and income received from foreign government pensions.

Receipts not counted as income include capital gains received (or losses incurred) from the sale of property, including stocks, bonds, a house, or a car (unless the person was engaged in the business of selling such property, in which case the net proceeds would be counted as income from self-employment); withdrawals of bank deposits; money borrowed; tax refunds; gifts; and lump-sum inheritances or insurance payments.

## REFERENCES

- Betson, David M. 1996. "Is Everything Relative?" The Role of Equivalence Scales in Poverty Measurement." University of Notre Dame, March.
- Betson, David. 1995. "Effect of Home Ownership on Poverty Measurement." Unpublished paper, November.
- Citro, Constance F. and Graham Kalton (eds.). 1993. The Future of the Survey of Income and Program Participation. Washington, DC: National Academy Press.
- Citro, Constance F. and Robert T. Michael (eds.). 1995. Measuring Poverty: A New Approach. Washington, DC: National Academy Press.
- Coder, John and Lydia Scoon-Rogers. 1996. "Evaluating the Quality of Income Data Collected in the Annual Supplement to the March Current Population Survey and the Survey of Income and Program Participation." Working Paper, U.S. Bureau of the Census. July.
- Cutler, David M. and Lawrence F. Katz. 1991. "Macroeconomic Performance and the Disadvantaged." Brookings Papers on Economic Activity No. 2, pp. 1-74.
- Doyle, Pat. Forthcoming (a). "How Can We Deduct Something We Do Not Collect? The Case of Out-of-Pocket Medical Expenditures." U.S. Bureau of the Census.
- Doyle, Pat. Forthcoming (b). "Who's at Risk? Designing a Medical Care Risk Index." U.S. Bureau of the Census.
- Eller, T.J. and Mary Naifeh. Forthcoming. "Housing Subsidies: Effect of Estimates on Poverty." U.S. Bureau of the Census.
- Fisher, Gordon M. 1992. "The Development and History of the Poverty Thresholds." Social Security Bulletin vol. 55 No. 4 (Winter), pp. 3-14.
- Garner, Thesia L., Geoffrey Paulin, Stephanie Shipp, Kathleen Short, Charles Nelson. 1997. "Experimental Poverty Measurement for the 1990's." Prepared for the Allied Social Science Meetings, session sponsored by the Society of Government Economists, January 1997; and the SGE Session: Measures of Well-Being From the Consumer Expenditures Survey, January 1997.
- Kokoski, Mary, Patrick Cardiff, and Brent Moulton. 1992. "Interarea Price Indices for Consumer Goods and Services: An Hedonic Approach Using CPI Data." U.S. Bureau of Labor Statistics, January.
- Moulton, Brent R. 1992. "Interarea Indexes of the Cost of Shelter Using Hedonic Quality Adjustment Techniques." U.S. Bureau of Labor Statistics, October.
- Orshansky, Mollie. 1963. "Children of the Poor." Social Security Bulletin v. 26 (July), pp. 3-13.
- Orshansky, Mollie. 1965. "Counting the Poor." Social Security Bulletin v. 28 (January), pp. 3-29.
- Ruggles, Patricia. 1990. Drawing the Line. Washington, D.C.: Urban Institute Press.
- Short, Kathleen, Martina Shea, and T.J. Eller. 1996. "Work-Related Expenditures in a New Measure of Poverty." Prepared for the 1996 Meetings of the American Statistical Association.
- Slesnick, Daniel T. 1992. "Gaining Ground: Poverty in the Postwar United States." Journal of Political Economy Vol. 101 No. 1 (February), pp. 1-38.
- U.S. Bureau of the Census. Money Income in the United States: 1995. Current Population Reports P60-193, Washington, DC: U.S. Government Printing Office, September 1996[a].
- U.S. Bureau of the Census. Poverty in the United States: 1995. Current Population Reports P60-194, Washington, DC: US Government Printing Office, September 1996[b].
- Watts, Harold W. 1993. "A Review of Alternative Budget-Based Expenditure Norms." Prepared for the Panel on Poverty Measurement and Family Assistance of the Committee on National Statistics, revised (May).
- Welniak, Edward J., Jr. 1990. "Effects of the March Current Population Survey's New Processing System on Estimates of Income and Poverty." Prepared for American Statistical Association annual meeting, August.

# VALUATION OF NONCASH BENEFITS

Table 12. **Income Distribution Measures by Definition of Income: 1995**

(Numbers in thousands. Households as of March of the following year. For meaning of symbols, see text)

Characteristic	Money income—			Before taxes			After taxes		
	Excluding capital gains (current official measure)	Definition 1 less taxes plus capital gains (losses)		Money income—		Definition 3 plus health insurance supplements to wage or salary income	Definition 4 less Social Security payroll taxes	Definition 5 less Federal income taxes	Definition 6 plus Earned Income Tax Credit
		Without EITC	With EITC	Definition 1 less government transfers	Definition 2 plus capital gains (losses)				
	1	1a	1b	2	3	4	5	6	7
ALL HOUSEHOLDS									
Total .....	99 627	99 627	99 627	99 627	99 627	99 627	99 627	99 627	99 627
Reciprocity Status									
With income as defined .....	99 032	99 032	99 032	93 004	93 009	93 009	93 009	93 014	93 014
With addition or deduction .....	(X)	(X)	(X)	42 392	15 918	54 312	75 096	73 158	14 860
Mean addition or deduction .....	dollars..	(X)	(X)	8 879	8 512	3 897	3 193	7 719	1 250
Standard error .....	dollars..	(X)	(X)	51	308	14	13	39	12
Mean total income .....	dollars..	(X)	(X)	23 715	85 353	64 598	51 682	47 964	20 696
Standard error .....	dollars..	(X)	(X)	269	1 309	419	343	252	232
Income Levels									
Percent .....	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Under \$5,000 .....	3.7	3.9	3.7	16.5	16.5	16.4	16.7	16.8	16.4
\$5,000 to \$9,999 .....	8.6	9.3	8.8	6.4	6.3	6.1	6.5	6.9	6.6
\$10,000 to \$14,999 .....	8.7	10.1	9.6	6.5	6.4	6.0	6.5	7.0	6.7
\$15,000 to \$19,999 .....	8.3	9.9	10.4	6.5	6.5	6.1	6.5	7.2	7.6
\$20,000 to \$24,999 .....	7.6	9.4	9.7	6.5	6.6	6.1	6.4	7.3	7.5
\$25,000 to \$29,999 .....	7.4	9.0	9.1	6.2	6.2	5.9	6.3	7.1	7.3
\$30,000 to \$34,999 .....	6.8	7.9	8.0	6.0	6.1	5.8	5.9	6.5	6.5
\$35,000 to \$39,999 .....	6.3	7.2	7.3	5.6	5.5	5.3	5.5	5.9	5.9
\$40,000 to \$44,999 .....	5.6	6.2	6.2	5.1	5.1	4.9	5.0	5.3	5.3
\$45,000 to \$49,999 .....	5.0	4.8	4.8	4.5	4.4	4.4	4.5	4.8	4.9
\$50,000 to \$59,999 .....	8.3	8.0	8.1	7.8	7.7	8.0	7.7	7.8	7.8
\$60,000 to \$74,999 .....	8.8	6.7	6.7	8.2	8.2	8.6	8.1	7.8	7.8
\$75,000 to \$99,999 .....	7.7	4.3	4.3	7.3	7.4	8.3	7.4	5.3	5.3
\$100,000 and over .....	7.1	3.4	3.4	6.8	7.1	8.1	6.9	4.4	4.4
Summary Measures									
Median .....	dollars..	34 076	29 093	29 219	30 931	31 082	32 819	30 793	28 393
Standard error .....	dollars..	197	135	134	166	171	215	193	173
Mean .....	dollars..	44 938	36 729	36 915	41 160	42 520	44 644	42 238	36 569
Standard error .....	dollars..	246	181	181	251	279	286	267	206
Income ratio .....	dollars..	444	418	414	503	511	509	514	490
Standard error .....	dollars..	.0039	.0039	.0039	.0038	.0040	.0039	.0040	.0039
Quintile Measures									
Lowest quintile									
Upper limit .....	dollars..	14 420	13 408	13 921	7 654	7 679	7 851	7 410	7 756
Percent of households .....	dollars..	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
With type of addition or deduction .....	dollars..	(X)	(X)	(X)	17 144	697	412	4 814	2 794
Mean amount .....	dollars..	(X)	(X)	(X)	9 666	-110	1 386	314	443
Standard error .....	dollars..	(X)	(X)	(X)	73	112	70	5	142
Second quintile									
Upper limit .....	dollars..	26 966	23 610	23 831	22 950	23 086	24 400	22 891	21 450
Percent of households .....	dollars..	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
With type of addition or deduction .....	dollars..	(X)	(X)	(X)	10 031	1 653	6 299	15 137	13 583
Mean amount .....	dollars..	(X)	(X)	(X)	9 354	796	2 054	1 017	1 550
Standard error .....	dollars..	(X)	(X)	(X)	102	89	22	8	9
Third quintile									
Upper limit .....	dollars..	42 012	35 288	35 397	39 659	39 940	42 235	39 619	36 021
Percent of households .....	dollars..	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
With type of addition or deduction .....	dollars..	(X)	(X)	(X)	6 685	2 489	13 412	17 691	19 353
Mean amount .....	dollars..	(X)	(X)	(X)	7 806	1 258	2 807	2 292	2 528
Standard error .....	dollars..	(X)	(X)	(X)	130	89	17	10	14
Fourth quintile									
Upper limit .....	dollars..	65 258	52 481	52 520	63 123	63 970	67 767	63 539	56 502
Percent of households .....	dollars..	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
With type of addition or deduction .....	dollars..	(X)	(X)	(X)	4 774	3 575	16 710	18 532	19 909
Mean amount .....	dollars..	(X)	(X)	(X)	7 189	2 310	3 877	3 566	5 187
Standard error .....	dollars..	(X)	(X)	(X)	165	98	20	14	23
Fifth quintile									
Upper limit .....	dollars..	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
Percent of households .....	dollars..	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
With type of deduction .....	dollars..	(X)	(X)	(X)	3 758	7 505	17 479	18 923	19 890
Mean amount .....	dollars..	(X)	(X)	(X)	8 073	16 371	5 477	5 987	1 201
Standard error .....	dollars..	(X)	(X)	(X)	196	623	28	30	327

Table 12. Income Distribution Measures by Definition of Income: 1995—Con.

(Numbers in thousands. Households as of March of the following year. For meaning of symbols, see text)

Characteristic	After taxes—con.								
	Definition 7 less State income taxes	Definition 8 plus nonmeans- tested government cash transfers	Definition 9 plus medicare	Definition 10 plus regular school lunches	Definition 11 plus means-tested government cash transfers	Definition 12 plus medicaid	Definition 13 plus other means-tested government—		Definition 14 plus net imputed return on equity in own home
							Noncash transfers	Noncash transfers less medical programs	
	8	9	10	11	12	13	14	14a	15
<b>ALL HOUSEHOLDS</b>									
Total .....	99 627	99 627	99 627	99 627	99 627	99 627	99 627	99 627	99 627
<b>Reciprocity Status</b>									
With income as defined .....	93 022	97 510	97 629	97 646	99 041	99 041	99 224	99 224	99 419
With addition or deduction .....	64 827	37 786	23 259	12 663	8 306	10 207	15 750	30 101	65 139
Mean addition or deduction .....	2 296	8 930	5 004	88	4 690	2 796	1 876	4 815	3 370
Standard error .....	26	54	26	1	68	38	22	26	30
Mean total income .....	44 052	31 024	34 655	57 171	19 596	31 942	21 925	15 056	50 829
Standard error .....	245	232	298	655	403	454	206	342	256
<b>Income Levels</b>									
Percent .....	100 0	100 0	100 0	100 0	100 0	100 0	100 0	100 0	100 0
Under \$5,000 .....	16 4	6 0	5 8	5 8	3 6	3 6	2 7	2 7	2 2
\$5,000 to \$9,999 .....	6 7	7 6	6 4	6 4	7 4	7 4	6 4	7 8	5 6
\$10,000 to \$14,999 .....	6 9	8 6	7 1	7 1	7 4	7 2	7 6	9 9	7 3
\$15,000 to \$19,999 .....	7 8	9 1	8 9	8 9	9 1	8 9	9 2	9 6	8 6
\$20,000 to \$24,999 .....	7 9	8 8	9 0	9 0	9 1	9 2	9 4	9 2	9 1
\$25,000 to \$29,999 .....	7 5	8 6	8 7	8 7	8 9	8 9	8 9	8 8	8 7
\$30,000 to \$34,999 .....	6 7	7 6	8 1	8 1	8 2	8 2	8 3	7 7	8 1
\$35,000 to \$39,999 .....	6 1	6 9	7 4	7 4	7 6	7 7	7 8	7 7	7 8
\$40,000 to \$44,999 .....	5 4	6 1	6 5	6 5	6 5	6 6	6 6	6 2	6 9
\$45,000 to \$49,999 .....	4 8	5 3	5 7	5 7	5 8	5 8	5 8	5 3	6 1
\$50,000 to \$59,999 .....	7 9	8 5	8 8	8 8	8 9	9 0	9 0	8 6	9 4
\$60,000 to \$74,999 .....	7 3	7 9	8 2	8 2	8 2	8 3	8 3	7 9	9 1
\$75,000 to \$99,999 .....	4 7	5 1	5 3	5 3	5 4	5 4	5 4	5 2	6 4
\$100,000 and over .....	3 8	4 0	4 1	4 1	4 1	4 1	4 1	4 0	4 8
<b>Summary Measures</b>									
Median .....	27 772	30 892	32 549	32 563	32 761	33 149	33 306	31 280	35 259
Standard error .....	163	156	146	144	142	143	143	153	154
Mean .....	35 262	38 649	39 817	39 828	40 219	40 566	40 802	39 347	43 006
Standard error .....	192	188	188	188	187	186	186	187	190
Gini ratio .....	481	424	412	412	404	400	394	409	388
Standard error .....	.0038	.0039	.0038	.0038	.0038	.0038	.0038	.0039	.0038
<b>Quintile Measures</b>									
<b>Lowest quintile</b>									
Upper limit .....	7 700	13 765	15 382	15 384	15 855	16 219	16 758	14 816	17 933
Percent of households .....	20 0	20 0	20 0	20 0	20 0	20 0	20 0	20 0	20 0
With type of addition or deduction .....	2 323	10 441	4 785	354	4 823	2 776	7 014	8 110	7 327
Mean amount .....	98	6 802	2 015	81	4 161	1 222	2 244	2 705	1 885
Standard error .....	8	47	28	3	62	28	34	26	66
<b>Second quintile</b>									
Upper limit .....	21 354	24 957	26 564	26 570	26 837	27 195	27 429	25 434	29 127
Percent of households .....	20 0	20 0	20 0	20 0	20 0	20 0	20 0	20 0	20 0
With type of addition or deduction .....	13 247	9 233	6 282	1 215	1 536	2 675	4 439	8 870	10 540
Mean amount .....	389	9 588	4 630	81	4 994	2 703	1 721	5 099	2 448
Standard error .....	5	86	26	2	169	46	41	41	48
<b>Third quintile</b>									
Upper limit .....	35 008	37 682	38 937	38 950	39 096	39 410	39 537	37 948	41 760
Percent of households .....	20 0	20 0	20 0	20 0	20 0	20 0	20 0	20 0	20 0
With type of addition or deduction .....	15 857	7 446	5 238	2 528	1 000	2 029	2 624	6 081	13 685
Mean amount .....	1 014	9 584	6 259	85	5 268	3 620	1 447	5 965	2 725
Standard error .....	8	116	49	1	259	85	49	59	45
<b>Fourth quintile</b>									
Upper limit .....	54 274	56 093	56 986	57 002	57 110	57 330	57 363	56 239	60 300
Percent of households .....	20 0	20 0	20 0	20 0	20 0	20 0	20 0	20 0	20 0
With type of addition or deduction .....	16 649	5 828	3 645	3 953	548	1 397	1 301	3 900	15 898
Mean amount .....	1 960	9 517	6 486	90	6 382	4 015	1 421	5 912	3 150
Standard error .....	12	160	61	1	382	131	75	80	50
<b>Fifth quintile</b>									
Upper limit .....	75 701	20 0	20 0	20 0	20 0	20 0	20 0	20 0	20 0
Percent of households .....	20 0	20 0	20 0	20 0	20 0	20 0	20 0	20 0	20 0
With type of deduction .....	16 751	4 839	3 102	4 612	399	1 151	371	3 141	17 689
Mean amount .....	5 657	10 653	6 412	89	6 142	3 908	1 393	5 872	5 231
Standard error .....	87	255	72	1	484	172	133	87	82

# VALUATION OF NONCASH BENEFITS

Table 12. **Income Distribution Measures by Definition of Income: 1995—Con.**

(Numbers in thousands Households as of March of the following year. For meaning of symbols, see text)

Characteristic	Money income—			Before taxes			After taxes		
	Excluding capital gains (current official measure)	Definition 1 less taxes plus capital gains (losses)		Money income—		Definition 3 plus health insurance supplements to wage or salary income	Definition 4 less Social Security payroll taxes	Definition 5 less Federal income taxes	Definition 6 plus Earned Income Tax Credit
		Without EITC	With EITC	Definition 1 less government transfers	Definition 2 plus capital gains (losses)				
	1	1a	1b	2	3	4	5	6	7
HOUSEHOLDS WITH FEMALE HOUSEHOLDER, NO HUSBAND PRESENT, WITH RELATED CHILDREN UNDER 18									
Total .....	8 751	8 751	8 751	8 751	8 751	8 751	8 751	8 751	8 751
Reciprocity Status									
With income as defined .....	8 670	8 670	8 670	7 653	7 653	7 653	7 653	7 659	7 659
With addition or deduction .....	(X)	(X)	(X)	4 467	666	3 630	6 728	4 455	4 648
Mean addition or deduction .....	(X)	(X)	(X)	6 188	5 590	3 327	1 683	3 241	1 622
Standard error .....	(X)	(X)	(X)	117	1 254	42	27	258	20
Mean total income .....	(X)	(X)	(X)	13 513	59 529	39 149	26 601	34 049	20 493
Standard error .....	(X)	(X)	(X)	442	5 353	840	646	671	483
Income Levels									
Percent .....	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Under \$5,000 .....	10.8	11.5	10.3	27.2	27.2	27.0	27.9	27.9	25.9
\$5,000 to \$9,999 .....	17.0	17.8	15.5	10.4	10.3	9.7	10.2	10.3	8.8
\$10,000 to \$14,999 .....	14.8	16.5	14.7	11.1	11.1	10.2	10.7	11.0	10.5
\$15,000 to \$19,999 .....	11.5	12.8	15.0	9.2	9.3	8.8	9.0	9.7	10.9
\$20,000 to \$24,999 .....	9.8	9.8	11.3	8.6	8.3	7.9	8.1	9.1	9.8
\$25,000 to \$29,999 .....	7.6	8.3	8.8	7.1	7.0	7.1	7.4	7.7	8.9
\$30,000 to \$34,999 .....	6.6	7.0	7.2	6.2	6.4	6.5	6.4	6.7	6.7
\$35,000 to \$39,999 .....	6.1	4.2	4.7	5.3	5.0	5.1	4.7	4.6	4.9
\$40,000 to \$44,999 .....	3.7	3.7	3.9	3.5	3.7	3.8	4.0	3.4	3.6
\$45,000 to \$49,999 .....	3.0	2.3	2.5	2.5	2.5	3.4	2.5	2.2	2.4
\$50,000 to \$59,999 .....	4.2	2.9	3.0	3.7	3.5	3.9	3.8	3.6	3.7
\$60,000 to \$74,999 .....	2.9	1.5	1.5	2.7	2.7	3.3	2.5	1.9	1.9
\$75,000 to \$99,999 .....	1.5	1.0	1.0	1.3	1.6	1.9	1.7	1.1	1.1
\$100,000 and over .....	1.3	.7	.7	1.2	1.2	1.3	1.1	.8	.8
Summary Measures									
Median .....	17 936	16 600	18 039	15 584	15 651	16 783	15 693	15 400	17 191
Standard error .....	409	303	287	395	393	456	431	395	367
Mean .....	24 508	21 504	22 365	21 349	21 774	23 154	21 860	20 210	21 072
Standard error .....	466	363	362	473	534	549	534	417	416
Income ratio .....	454	433	415	525	532	532	534	516	496
Standard error .....	0134	0134	0132	0129	0135	0133	0135	0127	0127
Quantile Measures									
First quintile									
Upper limit .....	14 420	13 408	13 921	7 654	7 679	7 851	7 410	7 351	7 756
Percent of households .....	41.3	40.6	37.3	33.0	33.1	32.7	33.1	32.7	30.7
With type of addition or deduction .....	(X)	(X)	(X)	2 413	21	38	1 271	25	744
Mean amount .....	(X)	(X)	(X)	6 513	(B)	(B)	277	(B)	972
Standard error .....	(X)	(X)	(X)	141	(B)	(B)	9	(B)	31
Second quintile									
Upper limit .....	26 966	23 610	23 831	22 950	23 086	24 400	22 891	21 450	21 834
Percent of households .....	24.8	24.9	27.1	30.1	30.5	30.0	29.5	29.0	29.4
With type of addition or deduction .....	(X)	(X)	(X)	1 080	86	1 045	2 389	1 202	2 224
Mean amount .....	(X)	(X)	(X)	5 406	1 069	2 487	1 035	682	1 999
Standard error .....	(X)	(X)	(X)	269	386	54	13	24	26
Third quintile									
Upper limit .....	42 012	35 288	35 397	39 669	39 940	42 235	39 619	36 021	36 127
Percent of households .....	18.9	18.5	18.9	21.7	21.2	21.7	21.5	21.5	22.7
With type of addition or deduction .....	(X)	(X)	(X)	607	209	1 374	1 757	1 785	1 209
Mean amount .....	(X)	(X)	(X)	5 789	1 667	3 080	2 041	1 706	1 381
Standard error .....	(X)	(X)	(X)	297	303	47	24	33	43
Fourth quintile									
Upper limit .....	65 258	52 481	52 520	63 123	63 970	67 767	63 639	56 502	56 551
Percent of households .....	10.6	10.9	11.5	10.6	10.8	11.2	11.5	12.0	12.3
With type of addition or deduction .....	(X)	(X)	(X)	246	185	844	936	1 034	378
Mean amount .....	(X)	(X)	(X)	6 778	2 699	4 015	3 247	3 812	1 501
Standard error .....	(X)	(X)	(X)	577	342	84	55	81	75
Fifth quintile									
Upper limit .....	4.5	5.1	5.2	4.5	4.4	4.4	4.4	4.7	4.9
Percent of households .....	(X)	(X)	(X)	121	165	328	376	408	93
With type of deduction .....	(X)	(X)	(X)	7 501	16 944	5 390	5 005	16 231	1 453
Mean amount .....	(X)	(X)	(X)	1 006	4 794	213	172	2 600	165
Standard error .....	(X)	(X)	(X)						

Table 12. Income Distribution Measures by Definition of Income: 1995—Con.

(Numbers in thousands Households as of March of the following year For meaning of symbols, see text)

Characteristic	After taxes—con								Definition 14 plus net imputed return on equity in own home
	Definition 7 less State income taxes	Definition 8 plus nonmeans-tested government cash transfers	Definition 9 plus medicare	Definition 10 plus regular-price school lunches	Definition 11 plus means-tested government cash transfers	Definition 12 plus medicaid	Definition 13 plus other means-tested government—		
							Noncash transfers	Noncash transfers less medical programs	
	8	9	10	11	12	13	14	14a	15
HOUSEHOLDS WITH FEMALE HOUSEHOLDER, NO HUSBAND PRESENT, WITH RELATED CHILDREN UNDER 18									
Total .....	8 751	8 751	8 751	8 751	8 751	8 751	8 751	8 751	8 751
Reciprocity Status									
With income as defined .....	7 660	7 944	7 954	7 966	8 675	8 675	8 739	8 739	8 742
With addition or deduction .....	4 188	2 361	531	1 828	2 964	2 476	5 294	2 709	3 139
Mean addition or deduction .....	1 015	5 536	3 933	80	4 915	2 797	2 659	3 327	2 317
Standard error .....	81	171	160	1	102	81	46	86	125
Mean total income .....	30 900	24 111	34 364	37 830	14 355	25 399	19 893	12 885	36 398
Standard error .....	625	639	1 724	1 037	398	799	386	980	683
Income Levels									
Percent .....	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Under \$5,000 .....	25.9	21.4	21.3	21.3	10.2	10.0	3.5	3.5	3.2
\$5,000 to \$9,999 .....	9.0	10.0	9.8	9.9	15.0	13.6	10.9	11.2	10.5
\$10,000 to \$14,999 .....	10.5	10.7	10.6	10.6	12.7	12.0	14.8	18.0	14.6
\$15,000 to \$19,999 .....	11.3	11.8	11.8	11.8	13.6	13.0	15.8	16.0	15.4
\$20,000 to \$24,999 .....	10.5	10.4	10.6	10.5	11.1	12.2	13.4	12.7	13.2
\$25,000 to \$29,999 .....	8.6	8.7	8.4	8.4	8.8	9.1	10.6	10.0	10.5
\$30,000 to \$34,999 .....	6.8	7.6	7.7	7.7	7.8	8.1	7.8	7.9	7.3
\$35,000 to \$39,999 .....	4.8	4.9	5.0	5.1	5.7	5.9	6.5	5.6	7.3
\$40,000 to \$44,999 .....	3.4	3.8	3.6	3.6	3.7	4.1	4.4	4.0	4.4
\$45,000 to \$49,999 .....	2.6	3.0	3.3	3.3	3.4	3.4	3.4	3.0	3.5
\$50,000 to \$59,999 .....	3.4	3.7	3.8	3.8	3.9	4.1	4.3	4.1	4.8
\$60,000 to \$74,999 .....	1.7	2.0	2.1	2.1	2.1	2.3	2.3	2.1	2.6
\$75,000 to \$99,999 .....	1.0	1.1	1.2	1.2	1.3	1.4	1.5	1.2	1.8
\$100,000 and over .....	.7	.7	.7	.7	.7	.8	.8	.7	.8
Summary Measures									
Median .....	17 086	18 306	18 527	18 539	19 400	20 569	21 786	20 529	22 360
Standard error .....	357	342	337	336	312	329	285	299	300
Mean .....	20 587	22 081	22 319	22 336	24 000	24 792	26 400	25 370	27 231
Standard error .....	386	390	392	392	381	383	372	366	382
Gini ratio .....	491	470	470	470	421	413	367	370	368
Standard error .....	0125	0125	0125	0125	0130	0128	0129	0131	0129
Quintile Measures									
Lowest quintile									
Upper limit .....	7 700	13 785	15 382	15 384	15 855	16 219	16 758	14 816	17 933
Percent of households .....	30.6	39.3	42.7	42.7	40.5	38.8	35.6	31.8	37.4
With type of addition or deduction .....	154	874	143	177	2 104	876	2 686	648	559
Mean amount .....	68	3 889	1 467	81	4 570	1 433	3 154	1 614	1 029
Standard error .....	8	158	152	4	95	49	67	63	172
Second quintile									
Upper limit .....	21 354	24 957	26 564	26 570	26 837	27 195	27 429	25 434	29 127
Percent of households .....	29.1	25.0	24.2	24.1	25.8	26.5	28.4	30.4	28.1
With type of addition or deduction .....	1 292	557	99	433	485	830	1 572	1 135	764
Mean amount .....	249	5 117	3 932	76	5 629	2 917	2 281	2 962	1 518
Standard error .....	9	293	256	3	322	92	83	90	156
Third quintile									
Upper limit .....	35 008	37 682	38 937	38 950	39 096	39 410	39 537	37 948	41 760
Percent of households .....	22.9	19.2	17.6	17.6	17.8	18.2	19.8	20.8	18.1
With type of addition or deduction .....	1 502	483	113	580	229	428	659	529	843
Mean amount .....	660	5 964	4 624	77	4 814	3 847	1 951	4 965	2 185
Standard error .....	17	387	190	2	439	199	132	236	189
Fourth quintile									
Upper limit .....	54 274	56 093	56 986	57 002	57 110	57 330	57 363	56 239	60 300
Percent of households .....	12.5	11.4	10.6	10.6	11.4	11.4	11.9	11.8	11.3
With type of addition or deduction .....	694	291	94	401	89	220	308	245	650
Mean amount .....	1 366	7 988	5 398	85	8 353	4 615	1 866	4 843	2 733
Standard error .....	44	575	376	3	1 102	388	216	415	255
Fifth quintile									
Percent of households .....	4.9	5.0	4.9	4.9	5.1	5.2	5.3	5.2	5.1
With type of deduction .....	346	157	82	236	57	123	69	152	323
Mean amount .....	4 925	10 337	5 587	86	(B)	4 800	(B)	5 224	5 939
Standard error .....	900	1 207	444	4	(B)	832	(B)	521	787

# VALUATION OF NONCASH BENEFITS

Table 12. **Income Distribution Measures by Definition of Income: 1995—Con.**

Numbers in thousands Households as of March of the following year. For meaning of symbols, see text)

Characteristic	Money income—			Before taxes			After taxes		
	Excluding capital gains (current official measure)	Definition 1 less taxes plus capital gains (losses)		Money income—		Definition 3 plus health insurance supplements to wage or salary income	Definition 4 less Social Security payments taxes	Definition 5 less Federal income taxes	Definition 6 plus Earned Income Tax Credit
		Without EITC	With EITC	Definition 1 less government transfers	Definition 2 plus capital gains (losses)				
	1	1a	1b	2	3	4	5	6	7
HOUSEHOLDS WITH MEMBERS 65 YEARS OLD AND OVER									
Total .....	23 732	23 732	23 732	23 732	23 732	23 732	23 732	23 732	23 732
Income Status									
With income as defined .....	23 592	23 592	23 592	20 124	20 124	20 124	20 124	20 124	20 124
With addition or deduction .....	(X)	(X)	(X)	22 374	3 572	4 251	7 673	10 500	40 819
Mean addition or deduction .....	dollars..	(X)	(X)	11 414	6 168	3 100	2 225	6 116	7 062
Standard error .....	dollars..	(X)	(X)	64	483	50	41	224	417
Mean total income .....	dollars..	(X)	(X)	18 631	54 360	57 737	42 439	36 990	20 747
Standard error .....	dollars..	(X)	(X)	347	2 051	1 446	1 036	584	884
Income Levels									
Percent .....	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Under \$5,000 .....	3.3	3.3	3.3	40.7	40.7	40.5	40.9	40.9	40.8
\$5,000 to \$9,999 .....	16.7	16.7	16.7	12.6	12.6	12.5	12.6	13.2	13.3
\$10,000 to \$14,999 .....	16.6	16.6	16.6	8.9	8.7	8.6	8.7	9.1	9.1
\$15,000 to \$19,999 .....	13.0	13.4	13.4	7.0	6.9	6.7	6.9	7.5	7.5
\$20,000 to \$24,999 .....	9.5	10.2	10.2	5.6	5.6	5.6	5.6	6.1	6.2
\$25,000 to \$29,999 .....	8.1	8.8	8.8	4.1	4.0	4.0	4.1	4.0	4.1
\$30,000 to \$34,999 .....	6.1	6.7	6.7	3.3	3.3	3.3	3.4	3.8	3.8
\$35,000 to \$39,999 .....	4.9	5.4	5.4	3.0	3.0	3.0	2.9	2.7	2.7
\$40,000 to \$44,999 .....	3.9	3.8	3.8	2.2	2.1	2.2	2.1	2.0	2.0
\$45,000 to \$49,999 .....	2.8	2.6	2.7	1.8	1.8	1.9	1.9	1.7	1.7
\$50,000 to \$59,999 .....	4.4	4.3	4.3	2.8	2.7	2.8	2.6	2.6	2.6
\$60,000 to \$74,999 .....	4.1	3.3	3.3	2.6	2.8	2.8	2.7	2.4	2.4
\$75,000 to \$99,999 .....	3.3	2.7	2.7	2.5	2.6	2.8	2.6	2.0	2.0
\$100,000 and over .....	3.9	2.1	2.1	2.9	3.1	3.3	3.0	2.0	2.0
Summary Measures									
Median .....	dollars..	20 503	19 959	19 994	8 427	8 447	8 552	8 348	8 277
Standard error .....	dollars..	236	204	206	226	231	231	226	207
Mean .....	dollars..	30 934	27 745	27 777	20 173	21 101	21 656	20 937	18 264
Standard error .....	dollars..	369	287	287	365	408	416	404	307
Ratio .....	dollars..	470	436	435	655	664	665	664	639
Standard error .....	dollars..	.0087	.0084	.0084	.0088	.0091	.0090	.0091	.0088
Income Measures									
First quintile:									
Upper limit .....	dollars..	14 420	13 408	13 921	7 654	7 679	7 851	7 410	7 756
Percent of households .....		34.0	31.5	32.9	48.5	48.5	48.7	47.9	48.8
With type of addition or deduction .....	(X)	(X)	(X)	11 213	514	100	1 011	59	293
Mean amount .....	dollars..	(X)	(X)	10 747	90	1 355	290	(B)	316
Standard error .....	dollars..	(X)	(X)	84	128	152	11	(B)	36
Second quintile:									
Upper limit .....	dollars..	26 966	23 610	23 831	22 950	23 086	24 400	22 891	21 834
Percent of households .....		28.0	26.4	25.3	24.3	24.3	24.8	24.8	24.2
With type of addition or deduction .....	(X)	(X)	(X)	5 478	838	842	2 255	4 032	365
Mean amount .....	dollars..	(X)	(X)	12 286	1 196	1 861	971	854	870
Standard error .....	dollars..	(X)	(X)	126	104	59	21	16	73
Third quintile:									
Upper limit .....	dollars..	42 012	35 288	35 397	39 659	39 940	42 235	39 619	36 127
Percent of households .....		17.4	18.2	18.0	12.2	12.2	11.8	12.5	12.2
With type of addition or deduction .....	(X)	(X)	(X)	2 649	757	1 186	1 757	2 942	191
Mean amount .....	dollars..	(X)	(X)	11 657	2 032	2 411	1 916	2 893	880
Standard error .....	dollars..	(X)	(X)	206	173	55	39	40	98
Fourth quintile:									
Upper limit .....	dollars..	65 258	52 481	52 520	63 123	63 970	67 767	63 639	56 502
Percent of households .....		11.0	12.7	12.6	7.7	7.5	7.4	7.5	7.5
With type of addition or deduction .....	(X)	(X)	(X)	1 622	549	1 010	1 010	1 757	1 757
Mean amount .....	dollars..	(X)	(X)	11 547	3 524	3 171	2 965	6 226	1 190
Standard error .....	dollars..	(X)	(X)	258	256	80	64	101	140
Fifth quintile:									
Upper limit .....	dollars..	9 6	11.2	11.2	7.2	7.6	7.3	7.6	7.2
Percent of households .....		9.6	11.2	11.2	7.2	7.6	7.3	7.6	7.2
With type of addition or deduction .....	(X)	(X)	(X)	1 412	913	1 113	1 381	1 710	45
Mean amount .....	dollars..	(X)	(X)	12 728	19 171	4 863	5 403	24 155	(B)
Standard error .....	dollars..	(X)	(X)	319	1 713	125	134	1 153	(B)

Table 12. Income Distribution Measures by Definition of Income: 1995—Con.

(Numbers in thousands Households as of March of the following year For meaning of symbols, see text)

Characteristic	After taxes—con								
	Definition 7	Definition 8 plus nonmeans-tested government cash transfers	Definition 9 plus medicare	Definition 10 plus regular-price school lunches	Definition 11 plus means-tested government cash transfers	Definition 12 plus medicaid	Definition 13 plus other means-tested government—		Definition 14 plus net imputed return on equity in own home
	State income taxes						Noncash transfers	Noncash transfers less medical programs	
	8	9	10	11	12	13	14	14a	15
HOUSEHOLDS WITH MEMBERS 65 YEARS OLD AND OVER									
Total .....	23 732	23 732	23 732	23 732	23 732	23 732	23 732	23 732	23 732
Reciprocity Status									
With income as defined .....	20 126	23 426	23 501	23 504	23 596	23 596	23 626	23 626	23 701
With addition or deduction .....	10 540	22 030	20 707	459	1 838	2 354	2 617	20 748	18 737
Mean addition or deduction .....	1 559	11 268	5 063	79	3 894	2 140	1 491	5 296	4 636
Standard error .....	50	64	27	2	133	62	34	29	58
Mean total income .....	30 749	27 582	34 904	67 312	22 190	32 571	18 819	14 762	40 487
Standard error .....	507	286	313	3 685	824	913	553	409	376
Income Levels									
Percent .....	100 0	100 0	100 0	100 0	100 0	100 0	100 0	100 0	100 0
Under \$5,000 .....	40 9	5 3	5 0	5 0	3 3	3 3	2 9	2 9	1 8
\$5,000 to \$9,999 .....	13 4	15 5	11 1	11 1	11 9	11 7	10 9	15 4	8 1
\$10,000 to \$14,999 .....	9 3	16 3	10 0	10 0	10 2	10 2	10 7	17 9	10 1
\$15,000 to \$19,999 .....	7 6	13 2	12 2	12 2	12 4	12 3	12 9	13 5	11 1
\$20,000 to \$24,999 .....	6 2	9 7	10 6	10 6	10 6	10 5	10 5	9 9	10 5
\$25,000 to \$29,999 .....	4 4	8 6	9 0	9 0	9 1	9 2	9 2	8 7	9 0
\$30,000 to \$34,999 .....	3 6	6 5	8 6	8 6	8 8	8 6	8 6	6 6	8 2
\$35,000 to \$39,999 .....	2 7	5 2	7 2	7 2	7 4	7 5	7 5	5 3	7 9
\$40,000 to \$44,999 .....	1 8	4 1	5 3	5 3	5 3	5 5	5 5	4 1	6 8
\$45,000 to \$49,999 .....	1 6	2 6	4 2	4 2	4 2	4 2	4 3	2 7	5 6
\$50,000 to \$59,999 .....	2 7	4 3	5 5	5 5	5 6	5 6	5 6	4 4	6 7
\$60,000 to \$74,999 .....	2 2	3 5	4 8	4 8	4 8	4 9	4 9	3 5	6 0
\$75,000 to \$99,999 .....	1 8	3 0	3 7	3 7	3 7	3 7	3 7	3 0	4 6
\$100,000 and over .....	1 8	2 3	2 7	2 7	2 7	2 8	2 8	2 3	3 5
Summary Measures									
Median .....	8 214	19 897	25 556	25 556	25 828	26 035	26 106	20 205	29 611
Standard error .....	203	205	262	262	258	251	251	232	276
Mean .....	17 571	28 031	32 448	32 450	32 752	32 964	33 128	28 498	36 789
Standard error .....	268	295	305	305	304	305	304	294	319
Gini ratio .....	633	448	420	420	414	413	409	435	393
Standard error .....	0086	0084	0080	0080	0080	0080	0080	0084	0078
Quintile Measures									
Lowest quintile									
Upper limit .....	7 700	13 785	15 362	15 384	15 855	16 219	16 758	14 816	17 933
Percent of households .....	48 8	33 2	27 0	27 0	27 5	27 9	28 8	35 4	26 4
With type of addition or deduction .....	1 324	7 197	4 012	34	1 005	788	1 645	6 019	3 633
Mean amount .....	80	7 656	2 005	(B)	3 012	689	1 622	2 927	2 392
Standard error .....	3	49	28	(B)	115	29	40	30	86
Second quintile									
Upper limit .....	21 354	24 957	26 564	26 570	26 837	27 195	27 429	25 434	29 127
Percent of households .....	24 2	26 6	24 7	24 7	24 3	24 2	23 8	24 8	23 0
With type of addition or deduction .....	3 964	6 051	5 718	28	314	516	436	5 736	4 252
Mean amount .....	536	11 896	4 637	(B)	4 254	1 955	1 223	5 927	3 483
Standard error .....	7	84	27	(B)	311	57	71	45	68
Third quintile									
Upper limit .....	35 008	37 682	38 937	38 950	39 096	39 410	39 537	37 948	41 760
Percent of households .....	12 4	18 1	20 8	20 8	20 8	20 5	20 1	17 8	20 1
With type of addition or deduction .....	2 351	4 058	4 772	53	245	400	236	4 087	4 206
Mean amount .....	1 105	12 964	6 325	(B)	5 794	2 844	1 375	6 517	4 389
Standard error .....	23	131	51	(B)	590	114	145	62	80
Fourth quintile									
Upper limit .....	54 274	56 093	56 986	57 002	57 110	57 330	57 363	56 239	60 300
Percent of households .....	7 4	12 0	15 0	15 0	14 9	14 9	14 8	11 8	16 6
With type of addition or deduction .....	1 446	2 618	3 432	118	134	325	157	2 662	3 561
Mean amount .....	2 068	14 116	6 517	76	5 011	3 660	1 261	6 464	5 447
Standard error .....	46	230	66	4	498	212	151	83	120
Fifth quintile									
Percent of households .....	7 2	10 2	12 5	12 5	12 5	12 6	12 6	10 2	13 9
With type of deduction .....	1 456	2 106	2 772	226	139	325	82	2 243	3 085
Mean amount .....	6 459	14 993	6 330	33	5 021	3 565	1 276	6 426	8 267
Standard error .....	287	379	75	4	603	238	242	92	239

Weinberg, Daniel H. 1996. "Changing the Way the U.S. Measures Income and Poverty." Prepared for the Canberra Group on Income Statistics, December.

\* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9, 1997. Daniel H. Weinberg Chief, Housing and Household Economic Statistics Division and Charles T. Nelson Assistant Division Chief for Economic Characteristics Housing and Household Economic Statistics Division U.S. Bureau of the Census Washington, DC 20233-8500 USA May 1997 Phone: (301) 763-8550 Facsimile: (301) 763-8412. E-mail: Daniel.H.Weinberg@cmail.census.gov, Charles.T.Nelson@cmail.census.gov

<sup>1</sup> This paper is largely based on Weinberg (1996) and Garner et al. (1997).

<sup>2</sup> The history of income questions asked on the Current Population Survey is from Welniak (1990).

<sup>3</sup> The fungible approach for valuing medical coverage assigns income to the extent that having the insurance would free up resources that would have been spent on medical care. The estimated fungible value depends on family income, the cost of food and housing needs, and the market value of the medical benefits. If family income is not sufficient to cover the family's basic food and housing requirements, the fungible value methodology treats Medicare and Medicaid as having no income value. If family income exceeds the cost of food and housing requirements, the fungible value of Medicare and Medicaid is equal to the amount which exceeds the value assigned for food and housing requirements (up to the amount of the market value of an equivalent insurance policy — the total cost divided by the number of participants in each risk class).

<sup>4</sup> These tables also include three additional variants (denoted 1a, 1b, and 14a).

<sup>5</sup> See Fisher (1992) for more historical detail on the development of the poverty thresholds.

<sup>6</sup> Also critical to the definition of poverty is whether to use an absolute or relative measure. A relative measure sets the poverty standard at a fixed fraction, say 50 percent, of some measure of the population's well-being such as median family income. Thus, under a relative poverty measure, only if the incomes for the families at the bottom of the income distribution improve relative to the rest of the distribution would poverty decline. The alternate method of measuring poverty and the one currently in use in the U.S., at least in theory, is more or less an absolute measure. When constructing an absolute measure, one attempts to measure the minimal consumption levels of as many goods as possible. The cost of that consumption bundle is then increased to account for necessary goods not included by use of a "multiplier." Orshansky measured only the cost of

a minimally adequate diet. Other proposals have suggested adding shelter, clothing, and medical care to the list. We restrict the discussion here to absolute measures; most observers expect the U.S. poverty concept to retain this feature.

<sup>7</sup> Data are from the Compensation and Working Conditions Branch, U.S. Bureau of Labor Statistics. The 1966 percentage is not strictly comparable to the 1996 figure.

<sup>8</sup> Exceptions are wages and salaries (we suspect that respondents sometimes report net instead of gross earnings) and workers' compensation (payments for injuries on the job.) There are early indications that changes to the SIPP questionnaire in 1996 have ameliorated these problems.

<sup>9</sup> A National Academy of Sciences panel on the future of the SIPP recommended moving toward the use of the SIPP for official income and poverty measurement (Citro and Kalton, 1993).

<sup>10</sup> A full review of budget-based approaches is in Watts (1993).

<sup>11</sup> There is also an issue about whether to use the official CPI or an experimental CPI created to correct for inaccurate measurement of housing costs in the official CPI prior to 1983. The next CPI market basket revision is scheduled for 1998.

<sup>12</sup> This section drawn from Appendix A of U.S. Bureau of the Census, 1996a.

<sup>13</sup> In general, inventory changes are considered in determining net income from nonfarm self-employment: replies based on income tax returns or other official records do reflect inventory changes. However, when values of inventory changes are not reported, net income figures exclusive of inventory changes are accepted. The value of saleable merchandise consumed by the proprietors of retail stores is not included as part of net income.

<sup>14</sup> In determining farm self-employment incomes, inventory changes are usually considered in determining net income only when they were accounted for in replies based on income tax returns or other official records which reflect inventory changes; otherwise, inventory changes are not taken into account.

<sup>15</sup> Child support paid and other inter-household transfers should theoretically be subtracted from income to avoid double counting, but the data necessary to do so are not collected.

# Theoretical and Technical Solutions for Preservation of Electronic Records in Finland

## **Outsourcing as a temporary solution**

The National Archives of Finland faced a severe problem at the end of the 1980's. Electronic records had replaced traditional paper records in many cases. Though the National Archives preferred - and still prefers - paper records to electronic ones, it was impossible to generate hard copy records from electronic records in all cases.

The reason is simple. The capacity of a computer enables it to handle records which are unmanageable in paper or microfilm. It is obvious that appraisal can't be based on the physical format of a record, but on the information the record contains.

Our first attempt was to outsource the preservation function to in the main computer centre of Finland in 1989. Information was stored in 'traditional' way, by 1/2" magnetic tape in mainframe environment. It was, without a doubt, a safe solution. But outsourcing had its obvious disadvantages, too.

Preservation was costly. The charge was 50 FIM/reel each month, including backup. This meant that the cost of preservation was calculated in a different way for electronic than for paper records. This fact had an influence in appraisal practices whereby preserving data was considered as 'expensive' while paper was regarded as 'cheap'. It meant that a great amount of electronic records was not preserved because the solution was considered too expensive.

Access to electronic records was difficult. It was costly to load a tape and run queries in a mainframe. Mainframe serves well as a multi-user database server, but it is too robust a tool for archived material. The basic problem of archiving data is to provide access to a very large volume of very little used data. Only in very rare cases do you have more than one simultaneous user for the same file. You don't need to worry about the speed of transactions - so you don't need a mainframe.

One main point was whether it was wise for an organisation to outsource its key functions. It was clear, that the use of a computer centre had to be a temporary solution.

*by Matti Pulkkinen \**

## **Solutions in middle of an economic crisis**

In the beginning of the 1990's the economy of Finland suffered a major setback. In 1991, the gross domestic product diminished 7.1% and continued to diminish until 1994. Government consumption expenditure sank, too.

It was a crisis. For the National Archives it wasn't possible to invest in a tape repository, although the need for a preservation system of its own became evident. It was, also impossible to have more personnel.

That made us think about what we actually did need.

Who should be able to preserve more.. If we don't have data, then nobody asks for it, and we may think that nobody needs data archives. This circular argument was sometimes made, when we considered our task.

## **Safety is a major problem**

Secondly, we should do it in a safe way. Safety means safety in every respect, and it means both safety of material and safety of citizens, too.

When we think of the long-term preservation of a record, we must take into account the ageing of the material and other such threats. Also, we mustn't forget that the society tends to change as well.

Let us take an example. In the year 1897 a protocol of the Överstyrelse för pressärendena - the Supreme office of press affairs - written and then preserved. It is there, in the repository even today. In one hundred years it has 'lived' through a coup de état in 1899, a general strike and rioting in 1905, abolition of Överstyrelse för pressärendena and rioting in 1917, a civil war in 1918 and series of bombardements in 1939 - 1940 and in 1941 - 1944. It has been preserved through economic crises in 1973 - 74 and 1992 - 94.

If we think of the permanent preservation of electronic records, we must accept, that the future is uncertain. We hope, of course, that the next century is happier than current one, but we can't be sure of it.

In the preservation of electronic records you always need

some manpower and supplies. You can't just forget your disks or cassettes in a repository. There is a risk, because a severe economic crisis may make the performance of your routine duties impossible. If you run out of money, you may loose your data.

The nature of an electronic record makes its preservation during a severe crisis a difficult task. If we think of threats like a hostile occupation or a totalitarian coup de état, the main benefits of electronic records -, easy access and rapid altering - become threats. The horrors of Rwanda were enabled partially by misuse of census records. .

### **Safety by replication**

We selected a file in logical sense as the object of preservation. A physical media can't be preserved readable long term. File formats and physical formats become technically obsolete. A record is a semantically coherent set of information, a semantical entity, and not a physical one. It is, however, true that a record can't exist without some physical container for the information. As a corollary we can say, that you can use different types of containers for a record. This means in practical level that we have freedom to use any media we want to preserve a record as long the semantical integrity of a record can be guaranteed.

We can replicate electronic records. Indeed replication is a key technique in our concept of security. We always use two different media in preservation. We take a master copy in DAT and a backup in 8 mm, and if we want, a CD R-copy, too. DAT copy is preserved in National Archives, backup in high security deposit in another place.

The choice of storage media was dictated by the lack of money - we couldn't possibly construct a traditional repository for 1/2" tape. By selecting a more compact media, we lost perhaps some of proven security of traditional 1/2" tape. We can compensate for it by a faster cycle of recopying of records - once in five years. For data cassettes, it is sufficient to use a safe, secure storage vault as a repository. It is difficult to gain illegal access and it will keep the temperature and relative humidity on a stable level. Inside the vault the data cassettes can be stored in fireproof cupboards to secure them from fire, water and magnetic interference. There has been, alas, a difficulty with our use of cupboards. Relative humidity has been too high, probably because the insulation material contained wasn't perfectly dry before we started to use them.

However, these cupboards are easy to evacuate, if necessary. Changes of temperature are very slow inside the cupboard, and we can transport material for example in an open lorry in wintertime.

### **Downsizing with Unix**

As a computer environment we use IBM RS/6000 C20 machines. The operating system is AIX 4. There are 1/2"

tape-, QIC-, 8-mm Exabyte-, CD R- and 2 DAT devices in the same machine, and we are about to connect a 3480-device in this machine, too. For security reasons, this machine works as stand alone, but we can temporarily connect it to our network, if we want.

Unix has been an optimal platform for archival preservation. It contains powerful script language, c-compiler (unfortunately an option in AIX 4) and all the device drivers we need. AIX is easy to manage.

You can do a lot with Unix programs like cut and grep; you don't need to load your data for example in a relational database to query it.

The cost of investment for the construction of the vault and the purchase of storage cupboards and computer equipment is the equal of the cost of outsourcing the electronic records preservation function over a three-year period.

### **The value of records**

The values that inhere in basically any records are of two kinds: primary values for the originating agency itself and secondary values for other agencies and private users. Records are, of course, originally created to fulfil the first of these: they are created to accomplish those purposes for which an agency has been created - administrative, fiscal, legal and operating, if we are referring to public records. But in archival perspective, more important is the second meaning - records are, after all, preserved in an archival institution because they have values that will exist long after they cease to be of current use, and because their values will be for others than the creators.

These secondary values of records can be furthermore divided into two kinds:

1. The evidence they contain, i.e. evidential value, referring to the value that depends on the character and importance of the matter evidenced.
2. The information that they contain, i.e. informational value, which may relate, in a general way, either to persons, or things, or phenomena. In modern archival science the evidential value is regarded as the principal value of a record.

There has been, however, little or no discussion about the nature of these two values. A closer study of both these aspects of preservation is important, and especially when we are discussing the preservation of electronic records.

### **Informational value**

The semantics of informational value is quite clear. We can here adopt the so called 'the naive theory of truth' on the semantic issue. From an epistemological point of view this theory in itself is not sufficient, but altogether it is 'good

enough' to be used in logic. This theory of truth implies, that, for example, the sentence 'there is an A' is true only and only if 'A' is a really existing entity. If there is no entity called 'A' in the world, the sentence is obviously untrue.

According to what we said earlier, we can conclude, that when any record includes true and meaningful statements, it also contains informational value. So if a record states, that there exist - or has existed in the time that record in question was created - an 'A', and if this statement is true - it can be verified to be true -, the particular record in hand contains unquestionable informational value.

### **Evidential value and speech act theory**

We can't, however, adopt this theory of truth used above to the matter of the evidential value of a record. While a receipt can be genuine or not, but semantically it can't be said to be either true or untrue. In fact, the evidential value is always bound in the use of the particular record - the record does not have evidential value per se.

The semantics of evidential value has to be derived from so called 'speech act theory'. This theory states, that a performative act of speech is neither true nor untrue. A more coherent way to make the distinction is to use the terms successful and unsuccessful. A judge, for example, can impose a penalty on a criminal in the court. He has the authority to do so. In this context his sentence 'the court fires you the sum of 1000 pounds' is successful, when and if the judge follows formalities stated by the law. The judge can't use the power given by the law to him outside of the court. So his sentence of punishment will be unsuccessful when imposed outside of the court.

### **Some conclusions**

We can now argue that the very nature of evidential value is based on semantics like this. It implies that we can use binary logic as a tool of falsification of evidential value. If and when we accept the argument generally shaped in this paper, we can furthermore define some formal criteria for the evaluation of evidential value.

Some essential - although maybe not all - questions of evidential value can be reduced in the two following concepts of authenticity and integrity.

The first of these, the problem of authenticity, is one of the key problems concerning the preservation of electronic records. This question can't be solved with those techniques used in data transfer. For example public key encryption is highly software-bound and sometimes also hardware-dependent. In archival preservation we cannot and we should not rely on public key encryption techniques. The other key issue in preservation, i.e., integrity, can also be easily violated, for example by loosing referential integrity or transactional integrity.

So we can state, that - in the strict sense - evidential value is lost when authenticity or integrity is lost. Many preservation programs for electronic records seem to be unaware of these problems mentioned, and to which should be given more investigation.

\* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9,1997. Matti Pulkkinen, KANSALLISARKISTO.

---

# Tying Everything Together with a Relational Database

Rapid change in delivery method means that the information which is necessary to access data is also changing rapidly. The density at which a tape or cartridge was written is critical information for reading the data on it while if the data is online the location is critical. On way of handling such changes in required information is to store it in a relational database.

What exactly is meant by a relational database? Fully relational databases, satisfying all conditions set out in Codd's definition, may only exist in theory and would be more than needed for this discussion. Here what is required is first that there be some degree of normalization of the data. As an example some studies have only one dataset while others have many, a record that tried to anticipate how many datasets is not likely to be very practical so put the study information in a study record and the dataset information in a dataset record with one record for each dataset and a key to the study record. Then it is necessary that there be some way of accessing these records together so that it looks as if two (or more) separate records are really one. Structured Query Language (SQL) is the accepted (complete with an ISO standard) way of doing this for a relational database.

Using the example of studies and datasets in a simplified version, let's say we have two tables which is the accepted term in relation databases for the structure in which the records, called rows, are stored. The first table is called studies and has fields, columns in relational terms, study\_num and title and the second table is called datasets and has the columns study\_num, dataset\_num, and name. Now a very simple example to make sure we are all on the same page. You would like to have a list of all study titles and the names of the datasets for each study. You could issue the SQL command

```
select studies.title,
       datasets.name
  from studies, titles
   where studies.study_num
 = datasets.study_num;
```

This will give you a list of study titles and dataset names which the title repeated for each dataset. Since you did not

---

*by Pat Hildebrand \**

---

request study\_num or dataset\_num you will not get them back.

The experiences being talked about here used Relational Database Management Systems. These applications were started under Ingres and migrated to Oracle when the university wide choice of a database

system made the switch. When dealing just with data some of our researchers have used SQL under SAS. The focus here is not on the specific relational database but rather on the concepts so specifics should be taken as concrete examples rather than the only way of doing it.

Although querying the database directly using SQL is an option and it is possible to use SQL scripts in place of some of the things used here what will be talked about are Oracle Forms applications (developed with Oracle Developer/2000), perl scripts, and C programs. The C could probably be replaced with perl but it was a very ambitious undertaking written by the system administrator. When a perl script has to interact with the database it is currently oraperl which is perl4. The perl5 scripts use DBI/DBD::Oracle and will go into production when the C application has been successful tested out against a more recent version of Oracle.

Because they are probably only of interest to show the wide range of uses I will briefly touch on the UNIX administration applications. The C application (using Pro\*C precompiler to interface with Oracle) is a print accounting program that keeps track of printing on our UNIX cluster and our NT network. Users are given an allotment for printing each semester and must pay for additional printing. At the end of each semester a reconciliation is done from a cron job to zero out any allocated funds not used and put in the next semester's allotment. The cron job is a perl script. The application for providing additional funds (user paying, refund because of bad printing, etc.) is a Forms application.

The other administrative application is for keeping track of our users. Since not everyone on campus can have an account on our system this application must check with another database on campus to determine if a person is in the correct department and has the correct status (no undergraduates). This is accomplished by means of a

database link. I don't know terminology for other databases but with Oracle a database link is a means of accessing a table in another Oracle database as if it were a table in the local database. The printing is related to system users by a column indicating what allocation of print funds they get.

Data access is a bigger issue as while we provide computing to a limited group we provide data to the entire university. Since we require the use without an account on our system to show up in person and present then campus ID we originally developed applications under Ingres that were used on our UNIX system in character mode so that calling in from home did not present a problem. When the applications migrated to Oracle and developed under Developer/2000, we found that it didn't make sense to develop in character mode any longer. Even with PPP when people called in from home they were still using character mode as the campus software had vt220 emulation. The database had become even more a part of the application under UNIX as that is when we started putting data on line so that there was even more information that we were keeping track of although the user probably used less of it.

When we designed the database we had been using tapes and cartridges for obtaining the data and at first the data was still used on the mainframe where the access was via a tape job. The tables included one for tape labels which also kept track of the tapes that were used for backing up accounts, temporary use, blank tapes entered into the system but not yet used, etc. Another table contained tape information such as the density, the character set, and whether the tape was labeled. At first glance it might appear that these two tables are each one row per table but the labels table has text which could require more than one row. Other tables are for the individual files on the tape.

When we started putting data online we could redesign the tables dealing specifically with the tapes to include online information or use the fact that the database was relational and use an additional table for the online information. We choose the later since not all data was being placed online.

Data requests, information about what is online, and the library system for hard copy documentation all share tables about the studies as well as having their own tables. For data are current system is a mixture of perl scripts, Forms applications, and even some cgi scripts for web access to the information. The web access takes care of the university wide access to the information. People on campus but outside those who have accounts on our system still have to present an ID and request access in person but now they know if the data are on campus. If the data are on campus they are able to get some information such as the size of files that they are talking about before they make the request so that they can make

sure they have the space. We still get people who only want a few numbers, often a statistic that they would have to calculate from a very large dataset, thinking that everything will fit on a floppy that already has a number of files on it, but the web seems to have found users who are better prepared to make use of the data when they come over to request access.

The heart of our data system is a series of perl scripts. When new data are received information is entered into the database about the study and dataset numbers and the type of file (data, documentation, program, etc.). This is actually a Forms application. As the information is entered a todo file is written with information from this table and a number for later identifying the file internally. Please note that there is additional information entered form a master-detail relation in the form and programmatically.

Perl scripts to read tapes/cartridges or process files which have been ftped or are on other media such as CD-ROM use the todo files to determine if a file should be read and do the processing. The todo files are also generated for existing tape/cartridge files to be put online and moving some files but not all from a CD-ROM to disk so the issue of whether or not a file should be read is real. When processing is complete (some, unfortunately, is still manual checking) a file is written with information that should go into the database as well as information needed to move the file from the processing area to where it is accessible to the users.

The names of these files are placed in another file that is read by a nightly cron job. The actual moving of the data and recording of the information in the database is done by this script. Some of the other things done by the script are to check the type of file, find out from the database where that type of file should be moved, check that there is enough space for the file, if need be and there is one available set the database to use a new directory, check that there is not currently a file in the directory with that name, and send e-mail about problems.

The library application is able to tell whether we have any hard copy documentation for a specific study and whether it is on hand or checked out, check thing out, and check them back in. If someone already has a copy of a specific piece of documentation checked out they are not permitted to check out a second copy of the same thing. Also if someone has overdue documentation checked out the application will say what is overdue and no further check outs are permitted for that individual until the overdue documentation is returned and the situation cleared is some other way.

All of the user information for requesting data and checking out documentation is the same table so that

changes do not have to be entered in multiple locations. As much of the information as possible is look up information from other tables. This not only makes the entry simplified but it also makes for fewer errors' in addition to avoiding typos this avoids ambiguous entries such as "student".

There are a lot of relations that exist in the services that we provide. Using a database allows us to restrict who can do what at what time. Using a relational database for the necessary information has made for greater accuracy, a simplification of things since once we have entered an individual say for data we don't have to do turn around and enter them again for checking out the documentation, and the ability to do automate some of the work.

### References:

- Date, C.J. with Hugh Darwen. *A Guide to the SQL Standard*, 3rd Ed. Reading, MA: Addison-Wesley Publishing Co. 1993.
- Edelstein, Stephen. *Learning Oracle Forms 4.5: A Tutorial for Forms Designers*. New York, NY: Relational Business Systems. 1995.
- Gundavaram, Shishir. *CGI Programming on the World Wide Web*. Sebastopol, CA: O'Reilly & Associates, Inc. 1996.
- Musciano, Chuck & Bill Kennedy. *HTML The Definitive Guide*. Sebastopol, CA: O'Reilly & Associates, Inc. 1996.
- Wall, Larry, Tom Christainsen & Randal L. Schwartz. *Programming Perl*, 2nd Ed. Sebastopol, CA: O'Reilly & Associates, Inc. 1996.

<http://www.hermetica.com/tecnologia/DBI>

Oracle Developer/2000 Documentation

Forms Developer's Guide  
Release 4.5  
Part No. A32505-1

Forms Reference Manual, Volume 1  
Release 4.5  
Part No. A32509-1

Forms Reference Manual, Volume 2  
Release 4.5  
Part No. A32510-1

Forms Advanced Techniques  
Release 4.5  
Part No. A32506-1  
Forms Messages and Codes  
Release 4.5

Part No. A32508-1

\* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9, 1997. Pat Hildebrand, Social Science Computing, University of Pennsylvania. pat@ssc.upenn.edu

# Emerging Internet Image Archives Visualizing Biological Species and Medical Conditions

This paper discusses Internet sites which enable the user to see images of plants, animals, and human disease conditions on computer monitors. The organization of the electronic archives at these sites commonly mimics bioscience taxonomy. Among the entities experimenting with biological images for the Internet are international and government science agencies, academic and research institutions, businesses, computer labs, interest groups, and innovative individuals.

by Leslie A. Brownrigg \*

Digital visualization for the Internet is driven by evolving technology and legacy materials. Most images presented are raster bit-mapped files in the Graphics Interchange Format (GIF, extension: .gif) or use the Joint Photographic Experts Group compression (JPEG/.jpg). GIF and JPEG standards are incorporated in Hypertext Mark-Up Language (HTML) and in software for viewing, editing, and printing images. Platforms, operating systems, and graphics software can almost universally import and export files as GIF and JPEG often though the standard for transporting images (TIFF) between formats. Software "plug-ins" allow additional image file formats to be viewed.

The physical source media for the electronic images of biological subjects include specimen, film and digital photographs, slides, prints, drawings, x-rays, magnetic resonating images, electrocardiograms, and sonograms. Images displayed were originally digitized in dozens of raster and vector image file formats native to proprietary equipment and software used to capture, scan, or edit electronic images.

Human factors figures in visualization. Can people recognize the subject of the image? Recognition involves cognition, culturally determined perception, cultural preferences for style, and the visual experience and training (technical or aesthetic) of the viewer. Composition and clarity critical for human perception may be affected by the view, shape, and indication of scale, details, or color if critical for identification: what about the subject is displayed. Focus, size, tones, background or context, the placement of the subject, and the method used to create a digital image affect human perception: how technically the image is made. Images with high resolution can convey finely detailed content and color tone gradients, can be

magnified (zoomed up), and require a higher density of pixels: the differentiated units of equal size that make up an electronic image. High pixel density implies more bytes of data and larger files. Large files tax the capacity of Internet packet transmission, and serving and receiving computers. "Lossy" principle

compressions (including JPEG) reduce file byte size by literally "losing" visual bits, degrading images irretrievably. Sizes of raster files posted for viewing on the Internet to depict or identify biological species and human disease conditions were observed to range from around five kilobytes for highly compressed photo "thumbnails" and black and white line drawings to hundreds of thousands of kilobytes. The large size of electronic image files is one reason why more image files are posted for downloading than for immediate viewing on the Internet. File size influences the commonly made decision to locate image files in a terminal position, at the end of an archive in an auxiliary Computer Graphics Interface (CGI) bin or CD-ROM that exclusively warehouses image files. On home pages or directories, links to images may be embedded in items of an inventory list of subject or file names, or brief descriptive text, or lossy previews. (1) The terminal or attached position of image files fortuitously permits linking them to and from documents within and among sites on a unique file address.

If directory and file names are semantically meaningful in human language, then their content can be found more easily by Internet search engines which automatically check directory and file names. Using Latin or common names to label the directories and files where information or images about a species are located sets up a smooth interface to browsers. Files named "C/GIF/Birch.gif" and "flora/Asteraceae/Launeas/L-aborescens/Laborescense.l.jpeg" can be identified as images from their extensions, .gif and .jpeg. One file identifies the subject of the image using a common name, birch, while the other uses a scientific name, *L. aborescense*. Each Latin Genus and species name is short hand metadata for a unique taxon, defined and characterized in biological data bases and monographs. Bioscience ranks life forms— Kingdom, Phylum, Class, Order, Family, Genus, Species (and their subs and branches). The second file is nested in a cascade of directory names which mark the subject's botanical

classification. Several Internet sites have adapted the taxonomic levels of biological classifications as a structure for their archives of electronic materials about species, materials in hierarchies. Taxonomic levels that group similar life forms gave a jump start to the logical organization of elaborate archives of biological species.

Several Internet enterprises are vying to construct taxonomies or phylogenetic lists to describe and compile data for all the living species on earth—or in a large region. Some of these sites incorporate illustration. Others are choosing technologies incompatible with the special requirements of image files.

The organization of "The Tree of Life: A Phylogenetic Navigation Systems" is instructive. Tree of Life is a federated site distributed on 18 servers in three countries and is linked to additional servers where specialized sites note their on-line "place" in the tree. Tree's "page" on frogs (*Salientia*) resides on the University of Texas server along with "Herps of Texas." Adding a refereed description for a higher order "clade" or "terminal taxon" to the Tree is facilitated by the data entry program MacClade which can read from and to the widely used Phylogenetic Analysis Using Parsimony (PAUP) taxonomic program. Each basic "page" in the Tree of Life covering a taxon or more generalized life form is illustrated with scientific line drawings and compressed photo images. Some images hide the HTML indices to clade Latin names by which Tree's internal web crawler navigates among its "pages." Images can be attached or linked at any point to Tree documents and generated from Tree's random and searchable internal browser facility. Tree's photographic images range from 11,000 to 328,000+ bytes in size. (2) Tree's is a collaborative effort among entomologists at the universities that serve and host this federation. Its content is best developed for insects.

UNESCO's Man and the Biosphere (MAB), national government scientific agencies, and research or academic institutions also initiated on-line sites with comprehensive ambitions. In a complex of sites sponsored and affiliated with U.S. government agencies, attempts are underway to link sites and to standardize meta-data aspects for the inventory and description of biological species. Agencies in three cabinet-level departments of the United States federal government and collaborators are forging distributed Internet inventory data bases. Next include the US Information Center for the Environment (ICE), National Biological Information Infrastructure (NBII), and "ITIS"- the Inter-agency Taxonomic Information System. (3) Standards, data base layouts, and "meta maker" data entry programs being tested emphasize location and text data and cannot accommodate images. (4) Images on the sites affiliated with this complex are rare, and tend to adorn home pages, or lie in ad hoc galleries, or buried in menus. (5)

Bioscience requires at least one "voucher specimen" be held in a scientific repository such as a botanic garden, herbarium, zoo, museum, university, or research organization to establish and reference the existence and scientific name of a unique species. Many Internet sites experimenting with visualizing originate from these institutions that collect and classify. Some sites cling to traditional presentations of specimen, showing animals confined in aquaria or cages or dried plants laid out on conventional herbarium sheets (6). Others deploy new media to photograph living organisms in fields and forests (7) or through the microscope.

Many institutions post digital peeks into their formidable collections (8) and label images as copyright, hesitant to reveal their research patrimony absent "pay per view" or use fees. Proceeds from the market for digital visuals can potentially reimburse the high costs of processing, archiving, and serving image files. On-line mechanisms include limiting access (on passwords) to subscribers, collecting fees to view or to download files, and receiving a payment for each "hit" to the site as use royalties from subscription services (9) or from advertisers. Commercial tie-ins are possible. Several sites display biological visuals to catalog items for sale and businesses sponsor sites and popular federations to attract customers. Until recently, one herbarium's Internet presence was hosted and sponsored by a garden supply site. Two popular sites posting images of felines were sponsored respectively by a pet food company and an airline. Access to the bulk of images on-line that visualize human disease conditions is already structured primarily through subscription services. Access to electronic archives of medical imaging tends to be strictly privatized to the archive's donor/users. An off-line market for digital images exists in CD-ROMs, software, and print ads.

Some of the most visually and technically sophisticated sites featuring numerous images of biological species prepared for immediately viewing on the Internet were designed and initiated by experimenting individuals and computer teams. Henriette Kress issues directories full of image files depicting plants, parsimoniously linking each file name she lists directly to an image file. Kress' botanical site served from Helsinki, Finland (10) is relayed by "mirror" sites on three continents, a tribute to the site's outstanding content and popularity. Tim Knight developed and maintains sites dedicated to images and other electronic media to depict the living primates and designed Internet sites for several zoos and scientific associations. (11) Knight's sites are worth "visiting" for their design and the quality and economy of images. A computer technology group at a Texas university is posting images of crops and wildlife. (12) The Korean Research and Development Information Center's BIOINFO computer imaging project serves 2300+ optimized animal pictures in its on-line archive which can be searched using English vernacular

names in the Roman alphabet or Korean names and script. (13) Independently, the designers of these four sites pioneered similar technical solutions to problems that Internet visualization projects confront. The exact technology (equipment, software, file format, and post digitalization editing) they used differs. In common, the four sites post attractive color images that can be magnified at least once, in files sized in the modest 20-250 kilobyte range. In common, the four sites created the digital images displayed by scanning film photographs or slides. They select among their own and contributed photographs those that can be successfully scanned at quite low pixel densities (under 100 dots per inch), a technique that limits image file size at the moment of digital creation. Photographs suitable for this technical approach must be in-focus, high contrast, high content, and centered on the subject.

It is striking how many Internet images show the faces of animals and flowers of plants, suggesting face or flower are regarded as the most recognizable or visually interesting feature. Images posted to portray biological species are not standardized as to perspective or view whether the organism is large or microscopic.

By contrast, medical imaging by x-ray, magnetic resonance (MRIs), and other devices to diagnose or monitor patients' disease conditions have strictly standardized views. Protocols exist for positioning subjects in relation to the image capturing device to achieve the correct views. Resulting images require a specialized training to interpret. A principle of file compression can be applied to highly standardized views: to reduce file size by removing extraneous features located in predictable areas of the image rather than to "loss" randomized pixels throughout the image. Fixed view medical images (14) can be stored without the extraneous features, and templates of what typically appears in the removed sections can be reinserted to "decompress" the image for viewing.

Human disease conditions are not registered in a single universal taxonomic hierarchy like that for biological species. There is less consistency in approaches to organizing medical images on the Internet. Each disease considered unique will be identified by its own diagnostic code and by vernacular or scientific names. Diseases are grouped into areas of specialization recognized in the medical community, that is, grouped either as related to a particular organ or system of the body, or to a disease processes, or stage in the life cycle. Thus, medical images cluster on sites for medical specialties. Images of melanoma skin tumors (15) can be found in dermatology and cancer archives.

Additional organizing principles are applied. Medical images on the Internet (and private Intranets) are sometimes organized in archives containing a single

technological type of medical image — just CAT-scans, just MRI, although from different institutions and concerning different patients, perhaps because interpreting each type of image is still another medical specialization. Medical imaging is sometimes grouped across types by source (hospital, clinic, practice) and an identifier like date of deposit. Medical images are organized in the "case" of one patient, either walled off in a confidential archive or broadcast as a didactic history in one of the commercial medical teaching and reference Internet subscription services.

Electronic images depicting human medical conditions produced daily due to their role in diagnosis and treatment are lost to retrieval and research when buried in incompatibly indexed and organized archives.

General observations follow relevant to sites with extensive on-line archives that include image files. Site designs typically redress the intrinsically large size of image files by applying compressions and by exercising selectivity. Giving computer directories and files names that have meaning in natural language makes them easier to search. Archives organized by establishing layered classes that generically includes every particular thing in its class, are easy to navigate.

#### **End Notes (Internet addresses of sites):**

(1) *Examples of image to image directories are Missouri Botanic Garden's directory at*  
<http://www.mobot.org/mobot/research/gallery.lagallery2.html>

<http://www.turtlebackzoo.org/tbz25.htm>

<http://www.cbs.nl/temp/lmi/ff300mr.htm>

(2) *Tree of Life's format is explained at*  
<http://phylogeny.arizona.edu/tree/home.pages/intro.html>  
 and see

<http://phylogeny.arizona.edu/tree/eukaryotes/animals/chordata/osteostraci/osteostraci.gif>

<http://phylogeny.arizona.edu/tree/eukaryotes/animals/chordata/lchthystegalia.gif>

<http://phylogeny.arizona.edu/tree/eukaryotes/animals/arthropoda/hexapoda/coleoptera/coleoptera.html>.

(3) *For more information about this U.S. government Internet enterprise see*  
<http://www.nbs.gov/nbii>

<http://trident.fic.nrcs.usda.gov/tius/>

<http://ice.ucdavis.edu/>

[http://ice.ucdavis.edu/US\\_National\\_Park\\_Service/National\\_Park\\_Service.html](http://ice.ucdavis.edu/US_National_Park_Service/National_Park_Service.html)

[http://ice/about\\_NPS\\_databases.html](http://ice/about_NPS_databases.html)

<http://www.fs.fed.us/>

(4) <http://www.itis.usda.gov/images/twbscr.gif>

<http://www.nbs.gov/nbi/current.status.html>

<http://www.fw.vt.edu/fishex/macsis.html>

(5) <http://www.nbs.gov/features/photogallery>

<http://www.itis.usda.gov/images/twbscr>

(6) <http://www.inbio.ac.cr/tipos/herbario/Cgrayumii.jpg>

[http://www.mpiz-koeln.mpg.de/80/~stuecher/ross/potato/sucense\\_complete.jpg](http://www.mpiz-koeln.mpg.de/80/~stuecher/ross/potato/sucense_complete.jpg)

(7) <http://bluehen.ags.udel.edu/udgarden.html>

(8) <http://www.nfrcg.gov.htm>

<http://www.nmhn.cgi-in/wdb/fish/catalog/query/catgno==00334647>

(9) <http://members.aol.com/zooweb/pictures>

(10) <http://sunsite.unc.edu/herbmed/pictures>

<http://sunsite.sut.ac.jp/pub/acadmeic/medicinal/alternative-healthcare/herbal-medicine/unc.edu>

(11) <http://www.selu.com/~bio/PrimateGallery/primates/species.html/index.html>

<http://www.selu.com/~bio/>

(12) <http://www.ics.tamu.edu/FLORA/gallery.htm>

<http://leviathan.tamu.edu:7071/slides.ef/>

<http://straylight.tamu.edu/tamu/tracy/chklcon.html>

<http://ranch.tamu.edu/rlem/>

(13) <http://bioinfo.kordic.re.kr/animal/>

(14) [http://www.laurie.umdj.edu/database/CollectedApps.acgi\\$ImageList\\_jpeg?11843](http://www.laurie.umdj.edu/database/CollectedApps.acgi$ImageList_jpeg?11843)

(15) See the Dermatology Online Atlas from the Erlanger Image Database

[http://ultact/med/kli/derma/hiblddb/diagnose/dg\\_m.htm](http://ultact/med/kli/derma/hiblddb/diagnose/dg_m.htm)

*and see MD Challenger sample photo:*

<http://www.embbs.com/img/i0000012.jpg>

\* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9, 1997. Leslie A. Brownrigg, United States Bureau of the Census

# Information that Come as Images: Overview of Issues

## Abstract

Increasingly information is available as images: colour, black and white - from satellites, photography, scanning in the biomedical sciences or scanning of printed sources, like codebooks accompanying numeric data and historic records. These images have very specific computer formats but need to be easily embedded, identified, searched, transferred and browsed or reproduced to paper to make them useful as information carriers. Principal media for distribution and access are CDROM and Internet. Briefly formats, standards and applications are discussed to indicate how much actual progress is being made in putting image information into the hands of researchers and interested users.

For the last five years and not only in the social sciences but in many areas of research, in CD-ROM and Internet publishing and in library services, information is increasingly available as images. There are several reasons for this increase:

1. more applications that create images : data visualisation, analytic mapping, flow charting to represent questionnaire routing, GIS systems, convenient document delivery
2. more applications that need images as content matter: computer assisted learning, distance learning, multimedia
3. increased transformation feasibility from other media (like printed or hand written sources) to images: including image enhancement and manipulation to correct imperfect originals
4. new types of hardware that can produce images: digitising video sources , digital photocameras, digitising to images in biomedical research and health care, remote sensing, forensic applications
5. new possibilities for distribution of series of images: improving Internet transfer capacity, new multi-page image formats that fit better with Internet client - server approaches, easy creation of single or small series of CD-ROM's that can hold large numbers

of images

*by Repke de Vries \**

6. new approaches that explore the next step from preservation to presentation by bringing images into structures like SGML and HTML designs or by betting on new formats like PDF

7. better availability of applications that - also in the public domain - view, browse series, and print images, including reproducing to colour on a low end desktop with colour printing; new (PDF) reader software that can "search" images of textual information with help of shadow pages that hold text approximations of the original after "dirty OCR"

8. recognition of images as another electronic source of information that needs appropriate bibliographic description and identification

In summary: more and different types of information are provided as images and at the same time there has been a shift in emphasis from long term storage with exact replication of the original and efficient compression towards giving access : finding ways to locate, search, browse and navigate the information content of images.

In particular a format like TIFF represents that longer established practice of preservation : it handles virtually any image original - level of detail, colour schemes (including black and white) and the like - has different compression choices (including a lossy one), is computer platform independent, is recognised as a standard and by consequence a format of choice for scanning stations and almost any application that one way or the other converts, transfers, shows, does OCR or reproduces images to print. In addition to long term storage it is efficient in simple distribution: like FTP, CD-ROM's with series of TIFF images and document delivery of scanned journals. Internet and TIFF though proved difficult to match: the format has the multi-page feature but this allows by it self only a linear forward and backward browsing and TIFF cannot be transferred to Internet client software with features like an increasing level of detail or with "page on demand" . Always a complete (multi-page) TIFF with all original detail, has to arrive over the Net before the image information can be viewed or printed. A possibly

very time consuming exercise. <sup>1</sup>

In terms of file formats and compression schemes did the Internet and the relatively slow speed of parts of the electronic connection between user and provider, sparse improvements to bring down transfer time : JPEG for example reduces original image file size drastically - but is a lossy technique: after decompressing there are differences with the original ; Progressive JPEG and Interleaved GIF are Web developments that help transfer by first presenting a rough impression to the user and gradually completing the image. <sup>2</sup>

This development forces archival organisations that both have to preserve an image collection and are information provider with that very same collection, to keep their images in two formats and face a dilemma. A preservation choice should be TIFF, probably off-line on CD-ROM's. But a different format would be needed to suit Internet presentation and electronic publishing.

Though bringing down transfer time is needed where Internet speed is not ideal, the real issue of access is a different one:

1. finding approaches to bring images into a structure together with other information that can have arbitrary other formats, like ASCII text, raw data files or even sound samples when biological research has habitat photos of birds together with recordings of their calls and field notes. Such a structure would also allow browsing and navigating with hyper-links and not just linear as in the single multi-page image file

2. finding ways to examine the information contained in images: scanned text can have a rough searchable real text duplicate after "dirty OCR", other images (like photographs or historical sources) can have meta-data attached that allow for searching and locating. This meta-data would take the usual form of keywording, applying subject classifications or standard bibliographic reference.

The two approaches go particularly well together when they complement each other : access to an image collection or single image after searching meta-data, with the images subsequently structured in a way that puts them into context with other available information and permits easy walk-about and retrieval.

The main choices to realise this kind of access seem to be the following:

#### **For structure:**

1. SGML and HTML as SGML applied to the Internet

#### **Pro's:**

- both meta-data and any further information type besides images, can be brought into this structure and at the same time keep their own specific (technical) format;
- it has hyper-linking;
- it is a general standard with good availability of software to code into this structure and to decode it;

#### **Con's:**

- HTML Web browsers which are themselves in the public domain may need additional commercial software to view and manipulate particular image-formats : this forces the provider to choose a format that is either supported natively by the common Web browsers or for which the additionally needed image-viewer comes free;
- The same software arguing holds for SGML
- The Data Documentation Initiative is an SGML structure initiative that explicitly defines the possibility to bring needed information into SGML in whatever format it comes: also questionnaire pages scanned to images (URL: <http://www.icpsr.umich.edu/DDI/> and the May 1996 paper by K Rasmussen "Convergence of Meta Data. The Development of Standards for Social Science Data" (contact the author at [boye@get2net.dk](mailto:boye@get2net.dk)).

2. PDF with its additional hyper-linking and annotation features

#### **Pro's:**

- public domain availability of PDF reader software;
- the reader software has all navigating, browsing, searching and image-viewing together in one application ;
- it has hyper-linking

#### **Con's:**

- the structure is limited to images and text information (no raw data for example) that both first have to loose their original format and have to be imported into PDF with commercial software;
- while HTML and SGML know many applications that decode this structure and explore its information content, has complete PDF with mixed image/text content, searching and hyper-linking only the (free) Adobe reader software for technical access
- Adobe, the initiator of PDF, can be found at URL: <http://www.adobe.com/>
- "Internet publishing with Acrobat" by G Kent, published by Adobe Press (1996) discusses "...creating and integrating PDF files with HTML on the Internet..." and marks Adobe's move to adapt PDF to Internet presentation

3. A SGML and Internet - HTML approach supported by PDF, where PDF only holds the images but the (free)

Adobe reader software brings sophisticated viewing and manipulation, certainly after the recent adaptation of PDF to Internet requirements like "image on request" from a multi-page file.

- the PSID Internet site can serve as an example: URL: <http://www.isr.umich.edu/src/psid/pdf.html> It has to be noted that contrary to the above, the PSID site and many other providers use PDF only as a distribution format or document delivery mechanism, which was made attractive by Adobe putting PDF reader (and printing) software in the public domain. None of the expanded features of PDF are utilised.

#### **For searching information hidden in images:**

##### **1. Database approaches**

Several scenario's exist:

- descriptive information is indexed and searched, with the images as search result;
- WAIS makes this possible for the Internet outcomes of "dirty OCR" on images from scanned documentation, are indexed and searched; a further linking scheme has to connect with the right image free-text indexing and searching, where the retrieved text has graphic-annotation links to images;
- ISYS database software and the International Social Survey Program CDROM are an example. URL: <http://www.za.uni-koeln.de/data/en/issp/index.htm>

##### **Pro's :**

-the database or search engine approach quickly brings results, like any indexed free-text search

##### **Con's :**

for the same reason is precision in search outcome not always high ;  
additional effort is needed to create descriptive and image linking information, or with text from some source available at least the links to images ( like in codebook text and scanned questionnaire examples);

##### **2. Subject classifications and keywording**

- The CASS Question Bank on the World Wide Web illustrates this:  
- URL: <http://kennedy.soc.surrey.ac.uk/qb/Welcome.html>  
- It has in fact both different subject trees and a search engine.

##### **Pro's :**

a very precise search that takes time when browsing subject trees but can still be fast when assisted by a thesaurus approach

##### **Con's :**

Even more than in rough free-text indexing, is human effort needed to apply keywords or classify each set of images (like a questionnaire in the CASS example)

3. Applying the commercial Adobe Acrobat software possibilities to expand the PDF format with searchable text that can be produced with internal "dirty OCR". This search is linear and images that have a non-text content would obviously need other Acrobat means to pinpoint particular images.: annotations and hyper-linking. The hyper-linking to relevant images within the single PDF-file can interestingly enough be realised with the idea of a "clickable map": particular clickable areas of an image, like a scanned Table of Contents or an overview picture of the human body, can be hot-linked to where further image information starts. This can create a very intuitive guidance in searching.

- A very recent CDROM produced by the German Zentralarchiv and the Dutch Steinmetz Archive holding part of the Eurobarometer questionnaires in the original languages, uses this type of clickable hyper-linking
- ICPSR produces CD-ROM's that have the search facility in PDF after internal "dirty OCR": for example the CD0013 "Health and Well-Being of Older Adults" prepared by NACDA (URL: <http://www.icpsr.umich.edu/nacda>) This way both the original codebook page is available in PDF as image and for searching the same format holds a text equivalent where possible.

##### **Pro's :**

- creating the descriptive information to search on is part of the Adobe software  
- the idea of "click and go"

##### **Con's :**

- the search is slow ;
- non-text images cannot do with automated "dirty OCR" and need complicated, manually added hyper-linking and annotations ;
- clickable links need human effort to organise and implement ;
- the extended PDF format with the above mentioned features, probably has to be regarded proprietary and would always need the hitherto free Adobe reader software

#### **Conclusion.**

The newer, extended features of PDF are too recent to have had much evaluation. Many services and products have been realised in PDF but predominantly as carrier for document delivery and distribution - following the bringing in the public domain by Adobe of the PDF reader software.

SGML (HTML) seems to offer the better, more general

structure for access, which structure can also hold the descriptive information, keywords or meta-data that a search engine could take for indexing to help locate particular images. HTML Web applications have brought considerable illustration of giving access to information as images but not in any complicated way, other SGML examples like the DDI still have to make their point.

This overview makes a distinction between preservation and presentation and has focused on the latter. It is an exciting new area of services, still open ended in direction but moving forward quickly.

\* Paper presented at IASSIST/IFDO '97, Odense, Denmark, May 6-9, 1997. R de Vries, Steinmetz Archive, Netherlands  
Email: repke.de.vries@niwi.knaw.nl

<sup>1</sup> The Netscape Web site "Inline Plug Ins: Image Viewers" nevertheless feature several TIFF applications

<sup>2</sup> Also the PNG image format is an interesting new development: URL: <http://www.w3.org/pub/Graphics/PNG/>

---

# Categorizing Event Sequences Using Regular Expressions

*This is a reprint of a paper that first appeared in the JASSIST Quarterly Vol. 21:2 without tables. We apologize to the authors for this error of omission and reproduce the paper here in its entirety.*

*by Lisa Sanfilippo &  
John Van Voorhis\**

---

must carefully define what events represent the phenomena of interest.

The technique described in this paper was developed as an alternative to existing algorithms and allows researchers to identify sub-patterns of events within sequences at the start of their analysis,

based on theoretical or practical considerations. Because this technique operates on a single sequence at a time, it is faster than processes that require comparing many sequences to one another.

## Introduction

Researchers who work with large sequential datasets are often limited in the kinds of analytic strategies they can use because of the sheer size of the data. Automated techniques for analyzing sequences were developed in the 1960s by scientists studying DNA, RNA, and proteins. In a classic volume on sequence analysis, Sankoff and Kruskal (1983) demonstrated its potential application for subjects as diverse as bird songs and macromolecules. In other work, Andrew Abbott developed "Optimal Matching" for sequence analysis in the field of sociology.

In this paper, we describe a technique for analyzing sequences using Regular Expression Matching (REM). This technique allows researchers to examine patterns in longitudinal data by condensing sequences of events into smaller, more tractable units. We also briefly discuss the development of a database structure that facilitates this kind of analysis.

Although all sequence analyses compare linear arrangements of symbols, whether in human behavior or DNA, they differ in their assumptions about what makes two sequences similar or different. Sequences in their original form often contain too much detail for useful comparison, since the possible permutations of occurrences can be limitless. Therefore, in all cases researchers must create the rules that define sequence similarity for their analyses.

Methods for determining sequence similarity are often referred to as sequence-matching algorithms. These algorithms are mathematical, and compare sequences without reference to the semantic or theoretical structures that created them. When using such methods, researchers who wish to place their analyses in an appropriate context

## The Project

To illustrate REM, we will describe how we used it to analyze the sequence of events that led to a child's placement into foster care in three states: Illinois, Michigan, and Missouri. We were looking for systematic demographic and geographic differences among children that correlated with the events they experienced in the child welfare system. REM was developed to describe and compare the pathways the children took through this system. The data were derived from the administrative data systems of the Illinois Department of Children and Family Services, the Michigan Family Independence Agency, and the Missouri Department of Social Services.

## Preliminary Data Processing

We received two data extracts from each state: one covering investigations of child abuse and neglect in the Child Protection System (CPS), and the other, services such as foster care to children in the Child Welfare System (CWS).<sup>1</sup>

We began by creating a project database for each state with the same essential structure. Each state's database contained tables for CPS and CWS data and one table for demographic information on the children. Next, we created an event table that contained all of the administrative events for all of the children in each system. We then transformed each child's events into a sequence variable or "history." Finally, we used regular expression matching to formulate "careers" by reducing the history sequences. At each step in the process, we preserved enough information from the previous step to retain flexibility in the subsequent steps. As the categories became broader at each step, the comparability of the data across states increased.

### Creating the Event Table

In this analysis we focused on four key administrative events:

- (1) indicated investigation, an investigation in which credible evidence of abuse/neglect was found,
- (2) unfounded investigation, an investigation in which no credible evidence of abuse/neglect was found,
- (3) case opening, when a case was opened for child welfare services, and
- (4) placement, when a child was placed in a foster home or institution.

patterns between sequences would make them unsuitable for analyses in their present form, we had not foreseen the amount of variation in the length of the sequences. For example, examining the distribution of event sequences revealed that many children experienced only one event, while others experienced up to fifty. This wide variation in length made it difficult to make meaningful comparisons among cases and suggested that we needed a method that would not rely solely on whole-sequence comparison. Therefore, we focused our attention on identifying the sub-patterns which we had observed in the sequences.

Table 1. Example Event and History Codes

Event Code	History Code	Description
0	0	Birth
1	1	Indicated Investigation
3	3	Unfounded Investigation
9	9	Case Opening
15	B	Foster Home
17	C	Hospital Health Facility
23	C	Department of Mental Health Institution
6	6	Case Closing
18	D	Home of Parent (Ending Placement)

We created one record for every event a child experienced in either the CPS or the CWS. We then coded every record with a number denoting a particular event type (See "Event Codes" in Table 1). These records contained the child's ID, an event date, and an event type code (See Table 2).

### Creating the History Sequences

We transformed each child's event records into a single sequence of codes, since as separate records the table structure was not appropriate for sequence analysis. To make the programming and its interpretation easier, we used only single-character codes in the history sequence. Although each history code represented a single event, a given code value could represent more than one type of event (See History Codes" in Table 1).

We first reviewed a frequency distribution of the history sequences to identify the most common sequences and to see the repetition of patterns within and among sequences. This review also revealed data entry errors that we could correct or eliminate, such as children receiving services before their birth or children being born multiple times.

Although we had anticipated that the variation in the

Table 2. Example Event Data for One Child

ID	Event Code	Event Date
125	0	Feb 13, 1978
125	3	Dec 16, 1991
125	3	Dec 17, 1991
125	3	Jan 5, 1992
125	1	Feb 2, 1992
125	3	Apr 21, 1992
125	9	May 20, 1992
125	17	May 20, 1992
125	23	Jun 15, 1992
125	18	Apr 17, 1993
125	6	Jul 10, 1993

### Creating the Career Sequences

One goal of our research was to elucidate the connection between CPS investigations and a child's subsequent placement in foster care. We had three initial questions: (1) What sequences of investigations **never** resulted in a child welfare case opening and placement? (2) What

sequences of investigations resulted in the child's first placement? and (3) What sequences of events resulted in the child entering the system without an investigation?

Because of our extensive work with the Illinois data and our contact with all three states regarding current and past practices and policies, we had some knowledge of what the most common patterns of events might be.

The following examples illustrate how this prior knowledge provided us with clues about what patterns to focus our attention on:

- We understood that the number of investigations a child experienced was not a critical factor in the caseworker's decision to place the child in foster care. We knew that children with histories composed solely of unfounded investigations were almost never provided with services, despite repeated contact with the department. Therefore, we believed that the number of *indicated* investigations would predict placement better than the raw number of investigations.
- We knew that, in one state, caseworkers were reluctant to remove children from their homes after only one indicated investigation unless they were in imminent danger. Thus, we expected that a child with one indicated investigation would be less likely to be placed into foster care than a child who had two or more indicated investigations.
- In all three states, we knew it was possible for children to experience a case opening and placement without an investigation of abuse or neglect, but we had no information on the frequency of such occurrences.
- Our prior analyses of the foster care data indicated that once in foster care, a child could move between placements numerous times before being returned home. Although the placements could be of different types, the child was still living away from his or her parents. As a result, we chose to treat a series of placements without a return home as one career event.

### Regular Expression Matching

It became apparent in looking at the sub-patterns that they could be represented by regular expressions, a notation used widely in the computer science field for specifying and matching sequences.<sup>2</sup> (See Appendix.)

We created a file listing the regular expression patterns we had decided to analyze along with a "career" code for each pattern which is shown in Table 4. We grouped the patterns in passes because we knew that certain patterns occurred only at the very beginning of the history and we needed to control the generation of the matching program. The first pass was used to remove any events that occurred

Table 3. Example History Record

ID	History Sequence
125	0333139CCD8

before a child was born. Since we were especially interested in the first series of investigations, we created a pass that only matched to initial investigation sub-sequences. The last pass, which was applied repeatedly until the history was exhausted, contained all of the sub-patterns we were investigating. From this pattern file we generated a series of programs to transform the data.

We used the AWK programming language for both our program generator and the matching programs themselves. An AWK program is composed of a series of pattern and action pairs. It automatically reads through data files one line at a time, and each line is matched against the patterns in the order they are listed in the program. When a line contains data that matches one of the patterns, the action associated with that pattern is executed. The patterns may contain regular expressions, while the actions are written in a language similar to the C programming language.

In our project, the program generator read the pattern file containing the sub-patterns of interest to us and generated a series of programs that used those regular expression patterns to process the history data. Each program in the series corresponded to a particular pass in the pattern file. If a pattern matched to the beginning of a history sequence, the matching characters were removed and the career code for that pattern was appended to the career sequence. The child's id, history, and career were then passed to the next program for the next pass. The final program passed the data back to itself until the history sequence was empty or until a fixed number of passes had been run. If the history sequence was completely matched, a lower case 'x' was appended to the career to indicate completion. An upper case 'X' was appended if more history remained after the maximum pass limit had been reached.

### Analyzing the Career Sequences

Since our analysis was limited to examining the sub-patterns that led to a child's first placement, we did not analyze children's entire careers. Instead, we only analyzed the first four career events after a child's birth.

Because the REM approach simply recoded the original history sequences, it preserved the unit of analysis, thus allowing us to attach explanatory variables such as year of first entry into the system, sex, race, and region<sup>3</sup>. Once this information was stored in one file, we aggregated the data by creating a crosstabulation which contained frequencies for every combination of the career sequences and the explanatory variables. These files were relatively small

Table 4 History Sequence Sub-patterns Written in Regular Expressions

Pass	Regular Expression	Career Code	Event Sequence Description
1	.+0	a	A series of events before the birthdate
	0	b	Birthdate
2	3*13*11*	A	One or more indicated investigations in a series
	3*13*	B	One indicated investigation in a series
	3+	C	A series of unfounded investigations
3, 4, 5, 6 etc.	9	D	Case opening
	3*1[31]*	E	At least one indicated investigation in a series
	3+	C	A series of unfounded investigations
	6	F	Case closing
	[A-Z][A-Z]	G	A series of placements
	+		
	[A-Z]	H	One placement

(fewer than 1,000 records) allowing us to import them into a spreadsheet program for final analysis and presentation.

### Conclusion

The REM technique described in this paper departs from more common pattern matching methods in that it incorporates theory and practice into the actual matching process. Using this technique, researchers can test their assumptions about the structure of a sequence. It is an iterative technique that allows the analyst to explore patterns in the data and to compare them across populations simply and quickly. Because the process of developing the career file is split into several steps (i.e., creating the event table, creating the history sequences, and pattern matching), it provides many opportunities to check the data and to ensure that the processes are transforming the data correctly.

REM allows the researcher to take a very large dataset and to represent it in a much smaller form, while maintaining the critical details of event order and sequence. For example, in our Illinois database we began with an event file of over 5 million records. Transforming this file into history sequences, career sequences, and finally into a crosstabulation, decreased the size of the file by a factor of 5,000, making it significantly easier to work with.

The REM technique, as written in AWK, can save the researcher hours of processing time, in large part due to: 1) the way AWK reads data files (i.e., it automatically reads a file one record at a time) and 2) the minimal programming it requires. Performing the same analyses using a statistical software package would have required much more extensive programming and perhaps more important, would have restricted the kinds of questions we could have asked in exploring the original data.

Table 5. Example Career Sequence Output

Output Record			
	ID	History Sequence	Career
Before Processing	125	0333139CCD6	
After Pass 1	125	333139CCD6	b
After Pass 2	125	9CCD6	bB
After Pass 3	125	CCD6	bBD
After Pass 4	125	6	bBDG
After Pass 5	125		bBDGF
After Pass 6	125		bBDGFx

### Special Characters Used in Regular Expressions:

.	Any single character
[ ]	Any single character listed within the brackets
+	Any sequence of one or more of the preceding symbol
*	Any sequence of zero or more of the preceding symbol
	OR - match one of the expressions the bar separates
( )	Treat the characters in the parentheses as a single symbol

*Review of Sociology*, 21:93-113.

Abbott, Andrew & Alexandra Hryciak. 1990. "Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers." *American Journal of Sociology*, 96(1): 144-185.

Abbott, Andrew & John Forrest. 1986. "Optimal

Matching Methods for Historical Sequences." *Journal of Interdisciplinary History*, 16(3): 471-494.

Aho, Alfred V., Brian W. Kernighan, & Peter J. Weinberger. 1988. *The AWK Programming Language*. Reading: Addison-Wesley.

Aho, Alfred V., Jeffrey D. Ullman. 1979. *Principles of Compiler Design*. Reading: Addison-Wesley.

Forrest, John & Andrew Abbott. 1990. "The optimal Matching Method for Anthropological Data: An Introduction and Reliability Analysis." *Journal of Quantitative Anthropology* 2:151-170.

Friedl, Jeffrey E. F. 1997. *Mastering Regular Expressions*. Sebastopol: O'Reilly & Associates, Inc.

Sankoff, David & Joseph B. Kruskal eds. 1983. *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley.

### NOTES

<sup>1</sup>. Child Protection Systems: in Illinois, the Child Abuse and Neglect Tracking System; in Michigan, the Protective Services Management Information System; and in Missouri, the Child Abuse and Neglect Data System.

Child Welfare Services Systems: in Illinois, the Child and Youth Centered Information System; in Michigan, the Children's Services Management Information System; and in Missouri, the Alternative Care Tracking System.

<sup>2</sup>. Regular expressions (REs) can recognize patterns which are left linear. Patterns, such as a balanced sequence of parentheses, cannot be recognized by REs.

### Future Directions

Clearly, REM has a much wider application than what we have illustrated with our project. Our analysis did not utilize REM to its fullest potential. For example, instead of analyzing just the initial sequence of sub-patterns, REM could be used to analyze full careers. We could run a similar process against the career sequences to further shrink the number of categories.

Finally, we did not explore the sub-patterns in as much detail as we could have. For example, we included specific placement event types in our event table and history sequences but did not treat them as separate types. In the future, we can easily compare differences in children's histories following specific types of substitute care placements (e.g., home of a relative, private foster home, group home, etc.) based on this project's current database.

### APPENDIX

#### Regular Expressions

In general, a character in an AWK regular expression matches itself. Some characters with special meanings in our pattern file are listed below along with some examples of their use. See the references for more details.

### REFERENCES

Abbott, Andrew. 1995. Sequence Analysis. *Annual*

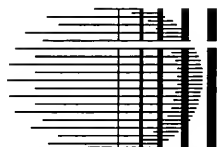
### Regular Expression Examples:

.	Matches 3, 6, 9, C, D, 6
[31]	Matches 3 or 1
3+	Matches 3, 33, 333, 3333, ...
3*1	Matches 1, 31, 331, ...
(31)+	Matches 31, 3131, 313131, ...
[31]+	Matches 3, 1, 331, 131, 333, 111, ...
(3 1)+	Matches 3, 33, 333, ..., and, 1, 11, 111, ...

because such patterns require "going-backwards" or maintaining information outside of the RE. For further information see Aho, Kernighan, and Weinberger 1988 in the references.

<sup>3</sup>. In all three states we differentiated the major urban area from the balance of the state.

<sup>4</sup> Paper presented at the IASSIST/IFDO 1997 Annual Conference Odense, Denmark May 7, 1997



# IASSIST

INTERNATIONAL ASSOCIATION FOR  
SOCIAL SCIENCE INFORMATION  
SERVICE AND TECHNOLOGY

• • • • •

ASSOCIATION INTERNATIONALE POUR  
LES SERVICES ET TECHNIQUES  
D'INFORMATION EN SCIENCES  
SOCIALES

## Membership form

The **International Association for Social Science Information Services and Technology (IASSIST)** is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data

Paid-up members enjoy voting rights and receive the **IASSIST QUARTERLY**. They also benefit from reduced fees for attendance at regional

and international conferences sponsored by IASSIST.

### Membership fees are:

Regular Membership: \$40.00  
per calendar year.  
Student Membership: \$20.00  
per calendar year.

Institutional subscriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subscription:  
\$70.00 per calendar year  
(includes one volume of the Quarterly)

I would like to become a member of IASSIST. Please see my choice below:

Options for payment in Canadian Dollars and by Major Credit Card are available. See the following web site for details:  
<http://data.lib.library.ualberta.ca/iassists/mbrship2.html>

- ☐ \$40 (US) Regular Member
- ☐ \$20 Student Member
- ☐ \$70 Subscription (payment must be made in US\$)
- ☐ List me in the membership directory
- ☐ Add me to the IASSIST listserv

Please make checks payable,  
*in US funds*, to IASSIST and  
Mail to:

IASSIST,  
Assistant Treasurer  
JoAnn Dionne  
50360 Warren Road  
Canton, MI 48187  
USA

Name: \_\_\_\_\_  
Job Title: \_\_\_\_\_  
Organization: \_\_\_\_\_  
Address: \_\_\_\_\_  
\_\_\_\_\_  
City: \_\_\_\_\_ State/Province: \_\_\_\_\_  
Postal Code: \_\_\_\_\_ Country: \_\_\_\_\_  
Phone: \_\_\_\_\_ FAX: \_\_\_\_\_  
E-mail: \_\_\_\_\_ URL: \_\_\_\_\_



